applied sciences

MDPI

Article

MDA: An Intelligent Medical Data Augmentation Scheme Based on Medical Knowledge Graph for Chinese Medical Tasks

Binbin Shi [1], Lijuan Zhang [2,*], Jie Huang [2], Huilin Zheng [1], Jian Wan [2] and Lei Zhang [1,2,*]

[1] School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
[2] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China
* Correspondence: 121107@zust.edu.cn (L.Z.); papaver_rhoeas@126.com (L.Z.)

Abstract: Text data augmentation is essential in the field of medicine for the tasks of natural language processing (NLP). However, most of the traditional text data augmentation focuses on the English datasets, and there is little research on the Chinese datasets to augment Chinese sentences. Nevertheless, the traditional text data augmentation ignores the semantics between words in sentences, besides, it has limitations in alleviating the problem of the diversity of augmented sentences. In this paper, a novel medical data augmentation (MDA) is proposed for NLP tasks, which combines the medical knowledge graph with text data augmentation to generate augmented data. Experiments on the named entity recognition task and relational classification task demonstrate that the MDA can significantly enhance the efficiency of the deep learning models compared to cases without augmentation.

Keywords: data augmentation; medical knowledge graph; natural language processing; language modeling; medical data augmentation

check for updates

Citation: Shi, B.; Zhang, L.; Huang, J.; Zheng, H.; Wan, J.; Zhang, L. MDA: An Intelligent Medical Data Augmentation Scheme Based on Medical Knowledge Graph for Chinese Medical Tasks. *Appl. Sci.* **2022**, *12*, 10655. https://doi.org/10.3390/app122010655

Academic Editor: Rafael Valencia-Garcia

Received: 24 September 2022
Accepted: 19 October 2022
Published: 21 October 2022

1. Introduction

Deep learning models are widely employed in natural language processing [1,2], image recognition [3], etc. In addition, the performance of deep learning models depends on the number of annotated datasets [4], especially in specific fields, deep learning models are more dependent on annotated datasets. Hence, few annotated datasets are applied for NLP in the medical field, which leads to the poor performance of many deep learning models [5]. With the development of the Internet of Things [6,7] and the soundness of intelligent medical systems [8,9], data augmentation that aims to generate a new dataset is proposed to enhance the accuracy of deep learning models in different tasks of natural language processing (NLP).

Automatic data augmentation is first utilized in computer vision to train more efficient models [10], especially for small datasets in different domains. However, in the medical field, most medical knowledge is recorded in text data, such as electronic medical records. Although data augmentation can augment complex texts, there are many problems in the field of Chinese medicine. Firstly, image data augmentation cannot be applied to natural language tasks because natural language is discrete [11]. Secondly, traditional text data augmentation ignores the semantic information in sentences, which leads to the degradation of the semantic stability of contexts and labels. Third, medical texts contain a large number of specialized vocabularies [12], and existing data augmentation cannot be exploited to enhance the specialized vocabularies. Therefore, we propose a novel data augmentation called medical data augmentation (MDA) that can effectively identify medical words and keep the semantics of the sentence the same.

In this paper, we propose MDA based on medical knowledge graph for Chinese medical texts. First, a large amount of medical knowledge is collected and stored as triplets to construct a medical knowledge graph, which can be utilized to identify medical keywords in sentences. Moreover, based on the medical knowledge graph, the MDA is applied to augment the text data and enhance the performance of deep learning models. In the end, the experimental results of the MDA are compared with different models on three downstream application tasks.

In summary, the contributions of our work can be summarized as follows:

1. We propose a medical data augmentation (MDA) method, which can effectively remedy the problem of semantic stability in the Chinese medical field;
2. We experiment on NLP with two tasks: Named entity recognition (NER) and relational classification (RC). The MDA outperforms the traditional text data augmentation in terms of F1-score, and the MDA can increase the diversity of the augmented data;

The rest of the paper is organized as follows: Section 2 introduces some related work on data augmentation. In Section 3, the MDA is described in detail. In Section 4, the performance of the MDA is evaluated on the four different datasets for the NER task and RC task, and several analyses are presented. Finally, we give some conclusions and discuss future research directions in Section 5.

2. Related Works

The existing deep learning models are based on a large amount of labeled data to train the tasks [13]. However, due to the small training datasets, deep learning models are often overfitted in a specific domain, which leads to the poor performance of these models [14]. Therefore, text data augmentation methods that can generate more samples are proposed to solve the problem of data scarcity in NLP tasks [15]. The existing text data augmentation methods can be divided into three types: paraphrasing augmentation, noising augmentation, and sampling augmentation [16].

The methods of paraphrasing generate augmented data with semantic differences from the original data. In addition, the augmented data carries semantic information that is very similar to the original data. In terms of the structure of sentences, the methods of paraphrasing create augmented data by reconstructing word paraphrases, phrase paraphrases, and sentence paraphrases respectively. In the methods of word paraphrases, Daval-Frerot et al. [17] proposed a thesaurus based on data augmentation method that utilized a universal thesaurus to classify synonyms of selected words, and then synonyms could be randomly replaced to generate augmented sentences. Due to the limitations of synonyms, Coulombe et al. [18] proposed a data augmentation method by replacing superposition words. In addition, they also integrated the features of the types of words that include adverbs, adjectives, nouns, and verbs. Although these data augmentation methods based on thesaurus can expand the sample, they do not increase the diversity of the sample. Hence, Xie et al. [19] exploited the English–French translation method that could perform back-translation on each sentence to increase the diversity of the augmented data. However, the quality of these augmented data and the problem of grammar are unsatisfactory. Therefore, Zhang et al. [20] introduced a discriminator to filter sentences in the translation model to enhance the quality of the augmented data. Moreover, Digamberrao et al. [21] presented language translation issues and effects on different languages. In addition, Perevalov et al. [22] and Bornea et al. [23] employed different languages corpus to construct a multilingual translation model, and the model could improve the accuracy of the augmented data. Furthermore, the quality of the augmented dataset can be further improved by exploiting the methods of phrase paraphrases. Hence, the method of word embedding [24] and the method of masking word filling [25] were proposed respectively, which could effectively generate new sentences through model training. To alleviate the problem of ambiguous grammar, the methods of sentence paraphrases were introduced. For example, Hou et al. [26] and LIet ai. [27] proposed a data augmentation model based on Seq2Seq to solve dialogue tasks in intelligent systems. However, the Seq2Seq model is difficult to capture aspect terms. Thus, a novel data augmentation model [28] was proposed, which combined the Seq2Seq and the transformer to reconstruct word fragments, and it achieved excellent performance in different scenarios.

Different from the methods of paraphrasing, the methods of noising often exploited the way of adding noise to disturb the original data [29], which cannot affect the original semantics. Yan et al. [30] proposed a new data augmentation method of sentence-level random swapping to classify datasets in the legal field. In addition, Longpre et al. [31] also used random swapping between sentences to contain complete semantics, but the generalization of the augmented sentence was not excellent. To this end, Yu et al. [32] and Xie et al. [33] introduced the deletion mechanism based on the attention mechanism and the mechanism combined with word-dropout, respectively. Among them, the deletion mechanism utilized a hierarchical attention network to obtain the attentions of the sentence, and the most important part of the sentence could be extracted by the attention to generate complex augmented sentences. However, the word-dropout mechanism was derived from the neural network language models, which could reduce the semantic information in the sentence to enhance the generalization of words. Moreover, the Mixup mechanism was proposed to alleviate the problem that the label data were changed by the methods of deletion. Especially, Guo et al. [34] proposed the wordMixup mechanism and the senMixup mechanism to generate augmented data between different labels. The first one performed sample feature fusion in Word embedding space, and the second one incorporated the feature of the hidden states in sentences. To contain the semantic information in a specific domain, Cheng et al. [35] applied two mechanisms to the adversarial augmentation method for machine translation tasks, which achieved excellent performance in Chinese–English, Anglo–French, and Anglo–German translation tasks.

Sampling augmentation is a data augmentation model for specific application scenarios, which combines some methods of paraphrasing with some methods of noising. Moreover, compared with paraphrasing augmentation and noising augmentation, sampling augmentation is difficult to train and has many limitations to the training datasets. Min et al. [36] utilized the method of subject/object inversion to augment the training datasets for the pre-training task, and the accuracy of the pre-training model was improved. In addition, Kang et al. [37] also combined the paraphrasing augmentation with rules to integrate the original data and augmented data for natural language inference. For different language models, the GPT-2 model [38], the masked language model [39,40], and the Sbert model [41] were introduced to reconstruct ambiguous sentences by fine-tuning the language model in different application scenarios. Moreover, Krill et al. [42] proposed a method based on media dynamic data to analyze the trend of COVID-19. Although sampling augmentation alleviates the problem of diversity, it cannot achieve the condition of training datasets for training in the specific field.

3. Our Proposed MDA

In this section, we present a description of the MDA that is motivated by the traditional text data augmentation, in which the semantics in sentences are not changed during the data augmentation. The MDA combines medical knowledge to enhance text data in different methods. To be specific, the datasets are constructed by the MDA in two modules: the medical knowledge graph module and the medical text data augmentation module. In the medical knowledge graph module, the medical knowledge graph is constructed based on the medical knowledge that is crawled from open-source websites. On the other hand, in the medical text data augmentation module, the datasets are transformed into different datasets at the word level. The specific block diagram of MDA is shown in Figure 1. Moreover, medical knowledge graph module and medical text data augmentation module are elaborated on the following part.

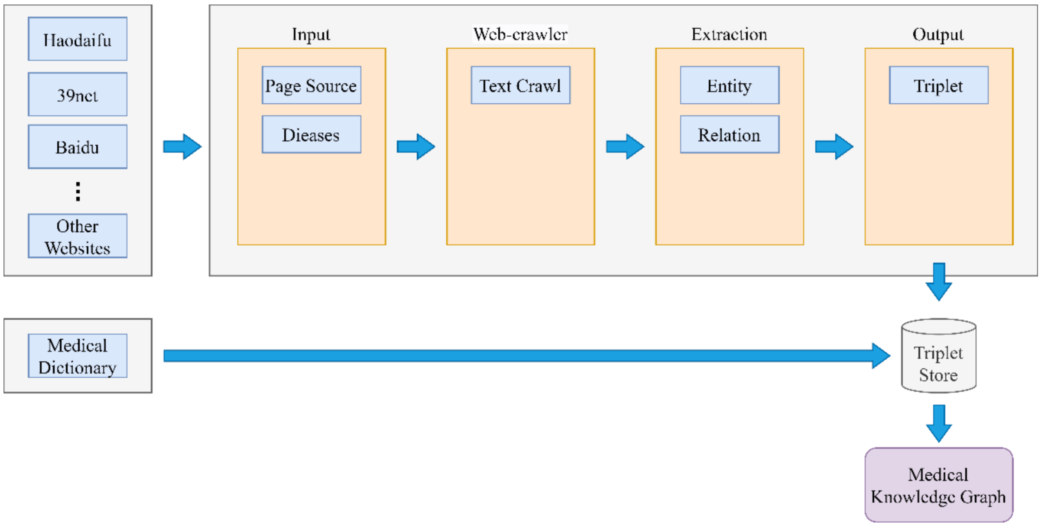

Figure 1. A framework of information extraction in the medical field.

Each element from the datasets is defined as **S**, and it consists of **x** and **y**, where **x** is a sentence that is made up of n words x_i, **y** is the label data of **x**, and the label y_i can be presented in different forms for different tasks. For example, in the NER task, each entity in the sentence **x** is tagged as a label y_i. In addition, the original set $\mathbf{S}_{ori} = \{\mathbf{x}_{ori}, \mathbf{y}_{ori}\}$ can be reconstructed to obtain the new set $\mathbf{S}_{aug} = \left\{\mathbf{x}_{aug}, \mathbf{y}_{aug}\right\}$ that the sentence and the labels are changed.

3.1. Medical Knowledge Graph Module

The articles of the open-source medical websites consist of structured information and semi-structured information, such as haodaifu, 39net, etc. Such information includes disease description, etiology, symptoms, complications, prevention, drugs, examination, and treatment methods. A medical extraction framework in the medical knowledge graph module is designed to extract the information from open-source medical websites and transform it into a medical knowledge graph composed of triplets. In this section, the detailed framework of information extraction in the medical field is shown in Figure 1, and it is structured into four parts: website source, web-crawl, extraction, and generation of triples.

First, many open-source medical websites are analyzed for the structure of the pages [43]. In addition, we propose different extraction schemas that depend on the composition of the content on the website. Second, medical knowledge from the website is crawled and stored as text, which consists of structured information and semi-structured information. Then, in the extraction parts, the structured information can directly utilize the method of segmentation and extraction to obtain medical knowledge. For example, the structured text 'drugs for cerebral infarction: recombinant tissue plasminogen activator, urokinase' ('脑梗塞的药品: 重组组织型纤溶酶原激活剂、尿激酶') is divided by punctuation to directly extract the drugs. However, for semi-structured information, the feature extraction model in NLP is exploited to recognize complex entities and relationships, such as NER model and RC model. After that, all the information that is extracted is stored as triplets in way of 'subject'—'predict'—'object'. Finally, the knowledge of medical dictionary includes nicknames for medical nouns, medical adjectives, and other supporting information, and it is combined with these triplets to construct medical knowledge graph.

3.2. Medical Text Data Augmentation Module

The medical text data augmentation module carries out different methods based on the medical knowledge graph to expand the original data, which can effectively increase and enhance the semantic stability of the data. Moreover, compared with the MDA, the

EDA [44] only focuses on the diversity of augmented data and ignores the relevance of words in sentences, and it includes synonym replacement (SR), random insertion (RI), random swap (RS), and random deletion (RD).

Table 1 shows a few examples of EDA from the original data: RS randomly selects consecutive words in the sentence and replaces them with words of the same type, such as heart disease and pneumonia; RI inserts adverbs or adjectives into sentences at random, such as words 'seriously'; RS randomly chooses two different words from the sentence and swaps their positions to generate augmented data; RD randomly removes any number of words from the sentence. Obviously, EDA does not require complex pre-training language models, and it is simpler than other data augmentation methods. However, the shortcomings of EDA are also obvious. First, because the original dataset **S** contains text data **x** and label data **y**, EDA only augments text data by keeping label data **y** consistency, but occasionally leads to text data to be incompatible with label data. Second, EDA is designed to generate large amounts of text data, which causes a sentence to abandon part of its semantics. In addition, the performance of the EDA is not effective for the text data of a specific domain.

Table 1. The EDA of four text data augmentation methods.

Methods	Examples	
	Original Data	**Augmented Data**
SR	He was sick with heart disease	He was sick with pneumonia
RI	He was sick with heart disease	He was seriously sick with heart disease
RS	He was sick with heart disease	He was disease with heart sick
RD	He was sick with heart disease	~~He~~ was sick with ~~heart~~ disease

To alleviate these problems, the medical text data augmentation module is designed with four methods that include medical knowledge replacement (MKR), words insertion (WI), words swap (WS), and words deletion (WD). To be specific, the medical text data augmentation module exploits the medical knowledge graph to identify medical terms and alleviate compatibility between texts and labels. Moreover, examples of augmented sentences are shown in Table 2.

In the MKR method, keywords are defined as words that are related to label data in text data. Each keyword in each label y_i is fed into the medical knowledge graph to obtain the medical knowledge that is associated with the keyword. Then, word that is related to the keyword is replaced with word from the acquired medical knowledge. It can be seen from Table 2 that an original data $\mathbf{S}_{ori} = \{\mathbf{x}_{ori}, \mathbf{y}_{ori}\}$ is given, where text data $\mathbf{x}_{ori}$ represents the text 'the patient with the cerebral hemorrhage had a headache' ('脑出血患者出现头痛') and label data $\mathbf{y}_{ori}$ represents the labels 'cerebral hemorrhage'-'symptoms'-'headache' ('脑出血-症状-头痛'). Obviously, entities 'cerebral hemorrhage' ('脑出血') and 'headache' ('头痛') can be considered keywords. After the keywords are input into the medical knowledge graph, a large amount of medical knowledge is obtained, such as 'cerebral hemorrhage'-'symptoms'-'dizzy' ('脑出血-症状-头晕'), 'cerebral hemorrhage'-'synonymy'-'cerebrovascular disease' ('脑出血-同义词-脑血管病'), etc. Hence, the augmented data $\mathbf{S}_{aug} = \left\{\mathbf{x}_{aug}, \mathbf{y}_{aug}\right\}$ can be generated, where text data $\mathbf{x}_{aug}$ represents the text 'the patient with the cerebrovascular disease had a dizzy' ('脑血管病患者出现头晕') and label data $\mathbf{y}_{aug}$ represents the labels 'cerebrovascular disease'-'symptoms'-'dizzy' ('脑血管病-症状-头晕').

Table 2. The MDA of four text data augmentation methods.

Methods	Examples			
	Original Data	**Keywords**	**Medical Knowledge**	**Augmented Data**
MKB	Text: 脑出血患者出现头痛 (The patient with the cerebral hemorrhage had a headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')	脑出血 cerebral hemorrhage 头痛 headache	脑出血-症状-头晕 ('cerebral hemorrhage'—'symptoms'—'dizzy') 脑出血—同义词—脑血管病等 ('cerebral hemorrhage'—'synonymy'—'cerebrovascular disease')	Text: 脑血管病患者出现头晕 (The patient with the cerebrovascular disease had a dizzy) Label: 脑血管病—症状—头晕 ('cerebrovascular disease'—'symptoms'—'dizzy')
WI	Text: 脑出血患者出现头痛 (The patient with the cerebral hemorrhage had a headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')	脑出血 cerebral hemorrhage 头痛 headache	头痛—形容词—严重的 ('headache'—'adjectives'—'severe') 头痛—形容词—复杂的 ('headache'—'adjectives'—'complex')	Text: 脑出血患者出现严重头痛 (The patient with the cerebral hemorrhage had a severe headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')
WS	Text: 脑出血患者出现头痛 (The patient with the cerebral hemorrhage had a headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')	脑出血 cerebral hemorrhage 头痛 headache	-	Text: 脑出血出现患者头疼 (The had with the cerebral hemorrhage patient a headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')
WD	Text: 脑出血患者出现头痛 (The patient with the cerebral hemorrhage had a headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')	脑出血 cerebral hemorrhage 头痛 headache	-	Text: 脑出血~~患者~~ 出现头痛 (~~The patient with~~ the cerebral hemorrhage had a headache) Label: 脑出血—症状—头痛 ('cerebral hemorrhage'—'symptoms'—'headache')

In the WI method, adverbs and adjectives of different diseases and symptoms are selected from the medical knowledge graph. In addition, different types of keywords such as disease and symptoms are obtained from the label data. After that, those selected words can be randomly inserted near the position of their associated keywords. It can be seen from Table 2 that the adverbs and adjectives of the keyword 'headache' ('头痛') in the medical knowledge graph include 'severe' ('严重的'), 'complex' ('复杂的'), and so on. Therefore, the augmented text data $\mathbf{x}_{aug}$ represents the text 'the patient with the cerebral hemorrhage had a severe headache' ('脑出血患者出现严重头痛'), and the label data $\mathbf{y}_{aug}$ is consistent with the original data.

In the WS method, two different words that are not associated with the label data in the sentence are chosen and the positions of the two are swapped. To be specific, the words that are not associated with the label data are 'patient' ('患者') and 'had' ('出现') in Table 2. When the positions of 'patient' ('患者') and 'had' ('出现') are swapped, the new text 'the had with the cerebral hemorrhage patient a headache' ('脑出血出现患者头痛') is obtained and the label data $\mathbf{y}_{aug}$ is consistent with the original data.

The WD method is similar to the WS method, the words that are not associated with the label data in the sentence are removed. For example, as shown in Table 2, after processing by the WD method, the new text 'the cerebral hemorrhage had a headache' ('脑出血出现头痛') is obtained and the label data $\mathbf{y}_{aug}$ is consistent with the original data.

Since medical records are mostly long texts, they can contain more label data and keywords. To judge the diversity of the augmented data, we define a parameter α that is the ratio of N to n, where N represents the number of changed words in the sentence, and n represents the length of the sentence.

In summary, the algorithm of the MDA is illustrated in Algorithm 1.

Algorithm 1: The Algorithm of the MDA

```
Input: Original dataset $\mathbf{S}_{ori}$, medical knowledge graph $\mathbf{G}$;
Output: Augmented dataset $\mathbf{S}_{aug}$, the probability of change $\alpha$;
for $i = 0$ to $|\mathbf{S}_{ori}|$ do
    Select $i$-th data $\mathbf{S}^i{}_{ori} = \{\mathbf{x}^i{}_{ori}, \mathbf{y}^i{}_{ori}\}$ from the original dataset $\mathbf{S}_{ori}$;
    Calculate the length $n$ of the text data $\mathbf{x}_{ori}$;
    Get the keywords $\mathbf{P}_{keys}$ from the label data $\mathbf{y}_{ori}$;
    Retrieve the medical knowledge in the medical knowledge graph $\mathbf{G}$ for the keywords
        $\mathbf{P}_{keys}$ to obtain all medical triplets;
    Use the methods of MKR, WI, WS, and WD sequentially to get the augmented data $\mathbf{S}_{aug}$,
        and the number of changed words in the MDA is $N$;
    Calculate the parameter $\alpha$ of the changed words in the original data $\mathbf{S}_{ori}$;
end
return $\mathbf{S}_{aug}$ and $\alpha$;
```

4. Performance Analysis

In this section, the experiments are introduced from four aspects. First, the data augmentation process is introduced. Second, three data sources and components of these datasets are explained. Third, deep learning models are described in detail, and these models are applied to NLP. Finally, the superiority of the MDA is verified by comparing the performance of the deep learning models in different tasks.

4.1. Data Augmentation Process

To compare the effects of the MDA, we first sample the original data to obtain the pending data. Then, with the support of the MDA, the pending data are transformed into augmented data, and the original data are combined with the augmented data to create the synthetic data. As shown in Figure 2, the deep learning models can be fine-tuned on the original data and the synthetic data for a specific task, respectively.

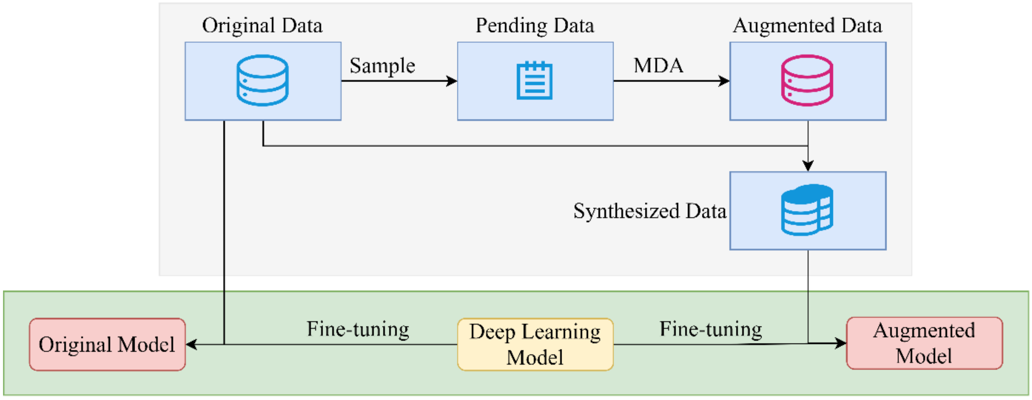

Figure 2. The prototype of the data augmentation process.

4.2. Datasets

The experiments are carried out on CCKS2019 dataset, CHIP2020 dataset, BITEmrNER dataset, and BITEmrRC dataset. In addition, we evaluate the MDA on the NER task and the RC task. Moreover, the structures of these datasets are described in detail as follows:

- CCKS2019 is a dataset for the Chinese electronic medical record NER task of the 13th China Conference on Knowledge Graph and Semantic Computing which aims to provide a platform for researchers and application developers to test technologies, algorithms, and systems, and it consists of 1379 medical records that include six categories of entities, namely anatomy, disease, imaging examination, laboratory examination, drug, and operation. During the fine-tuning process, CCKS2019 is divided into a training dataset and a testing dataset in a certain proportion;
- CHIP2020 is a Chinese medical dataset for the RC task of the 6th China Health Information Processing Conference which is an annual conference on biological information processing and data mining, which contains 43 categories of pre-specified relations, 17,000 Chinese medical sentences, and 50,000 triplets. Moreover, the dataset consists of 518 pediatric diseases and 109 common diseases. In addition, the CHIP2020 includes more than ten categories of entity, such as symptoms, imaging examination, etc. Moreover, to enhance the normalization of the dataset, the CHIP2020 is divided into a training set and a testing set in the official method;
- BITEmrNER is a Chinese medical dataset for the NER task collected by the BIT laboratory, and it consists of 1200 electronic medical records with cerebrovascular disease from the First Hospital of Zhejiang Province and the Fourth Affiliated Hospital Zhejiang University of Medicine. Furthermore, the BitEmrNER consists of the medical description text data and label data, where the text data include the history of present illness and past medical history, and the label data include different categories of entities. In this study, we randomly select 900 samples from the BITEmrNER to augment the training set;
- BITEmrRC is a Chinese medical dataset for the RC task collected by BIT laboratory, and it consists of electronic medical records with cerebrovascular disease from the First Hospital of Zhejiang Province and the Fourth Affiliated Hospital Zhejiang University of Medicine. In addition, the BITEmrRC contains more than 40 categories of pre-specified relations, 2400 Chinese medical sentences, and more than 8000 triplets. In this study, the BITEmrNER randomly selected 1600 samples from the BITEmrNER to augment the training set.

Therefore, it can be seen from Table 3 that the statistics of these datasets are listed to introduce the characteristics of CCKS2019 dataset, CHIP2020 dataset, BITEmrNER dataset, and BITEmrRC dataset. In fact, BITEmrNER and BITEmrRC datasets that are from the same set of electronic medical records and annotated for different tasks are electronic medical record datasets constructed by the BIT laboratory. However, the public datasets CCKS2019 and CHIP2020 pay attention to different diseases, and researchers can compare the performance of different models on the public dataset. In addition, the samples of these datasets are provided in Table 4.

After that, these datasets are employed in the different NLP tasks to verify the performance of MDA.

Table 3. Statistics of datasets.

Statistics	**CCKS2019**		**CHIP2020**		**BITEmrNER**		**BITEmrRC**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Sentence	1000	379		3585	900	300	1600	800
Task	NER		RC		NER		RC	
Categories	6		43		6		40	

Table 4. The samples of datasets.

Datasets	Text	Label
CCKS2019	患者4月余前因“下腹疼痛不适, 伴反酸、嗳气”在我院完善相关检查后确诊“胃体胃窦癌”。病人在全麻上行“远端胃大部分切除”’。在我院行SOX (奥沙利铂+替吉奥) 方案化疗。 (Before the end of 4 months, the patient was diagnosed as ‘gastric antrum cancer’ in our hospital due to ‘lower quadrant abdominal pain with acid regurgitation and belching’. The patient underwent ‘distal gastrectomy’ under general anesthesia. SOX (Oxaliplatin+ Tegafur) regimen chemotherapy was performed in our hospital.)	腹-解剖部位, 胃体胃窦癌—疾病和诊断, 远端胃大部分切除—手术, 奥沙利铂—药物, 替吉奥—药物 (‘abdominal’—‘anatomic site’, ‘gastric antrum cancer’—‘diseases and diagnoses’, ‘lower quadrant abdominal pain with acid regurgitation and belching’—‘operation’, ‘Oxaliplatin’—‘drug’, ‘Tegafur’—‘drug’,)
CHIP2020	按病程发展及主要临床表现, 可分为急性、慢性及晚期血吸虫病。急性血吸虫病多见于夏秋季, 以小儿及青壮年为多。 (Schistosomiasis can be divided into acute, chronic and advanced schistosomiasis according to its course and main clinical manifestations. Acute schistosomiasis is more common in summer and autumn, especially in children and young adults.)	血吸虫病—病理分型—急性血吸虫病, 急性血吸虫病—病理分型—吸虫病, 急性血吸虫病—多发群体—小儿及青壮年 (‘Schistosomiasis’—‘pathological classification’—‘acute schistosomiasis’, ‘acute schistosomiasis’—‘pathological classification’—‘fluke disease’, ‘acute schistosomiasis’—‘groups’—‘children and young adults’)
BITEmrNER	脑出血患者出现头痛, 呕吐。患者查头颅CT左侧额颞顶部硬膜下出血。 (The patient with intracerebral hemorrhage had headache, vomiting. The patient was examined on head CT for subdural hemorrhage on the left frontotemporal parietal side.)	脑出血—疾病, 头颅CT—影像学检查 (‘intracerebral hemorrhage’—‘disease’, ‘head CT’—‘imaging examination’)
BITEmrRC	脑出血患者出现头痛, 呕吐。患者查头颅CT左侧额颞顶部硬膜下出血。 (The patient with intracerebral hemorrhage had headache, vomiting. The patient was examined on head CT for subdural hemorrhage on the left frontotemporal parietal side.)	脑出血—临床表现—头痛, 脑出血—临床表现—呕吐, 脑出血—影像学检查—头颅CT (‘intracerebral hemorrhage’—‘clinical manifestation’—‘headache’, ‘intracerebral hemorrhage’—‘clinical manifestation’—‘vomiting’, ‘intracerebral hemorrhage’—‘imaging examination’—‘head CT’)

4.3. Results and Discussion

In this section, the results of the comparative experiments in NER task and RC task are analyzed and discussed. Next, the ablation experiments show the performance of each method in MDA. Finally, a sample is selected to analyze its original data and augmented data in the case study.

4.3.1. NER Task Evaluation

In order to evaluate the superiority of the MDA for the NER task, we choose two datasets to repeat the experiment five times. In the experiment, different data augmentation methods are exploited to generate the augmented data, then the original data is combined with the augmented data to build the synthetic data. Furthermore, three baseline models are utilized to recognize different entities. First, the model of BiLSTM-CRF is proposed by Huang et al. [45], where BiLSTM-CRF is proved to be a better NER model than other options such as BiLSTM, LSTM, CRF alone, or combinations such as LSTM-CRF. Second, BERT-CRF is utilized for the baseline in Chinese NER tasks by many researchers [46] because BERT is a widely used embedding model in many NLP tasks. Third, the Ra-RC model [47] combines radical features and a deep learning structure, and the performance of these models is evaluated based on the original data and the synthetic data.

For the deep learning models, the BERT model is employed to capture the features in sentences, and the CRF model is exploited to mark the recognized entities. In addition, the pre-trained weights that are obtained from the official release are not modified. Therefore, in this study, the batch size is set to 16, the learning rate is set to 1E-5, and the maximum length of the input sentence is set to 128.

Table 5 shows the results of the F1-score on the CCKS2020 dataset and the BITEmrNER dataset. Specifically, EDA and MDA are used to generate synthetic data from the original data, and the number of synthetic data is equal to the number of the pre-set. For the EDA settings, the four transitions (SR, RI, RS, RD) are randomly selected with a 40% replacement probability. For MDA Settings, the four methods (MKR, WI, WS, WD) are successively utilized to expand with a replacement probability of 40%. For the CCKS2019 dataset, it can be observed from Table 5 that the EDA outperforms the original data in the entity recognition by 1.04% and the MDA outperforms the original data by 2.77% in the best performance. The improvement of the EDA can be explained by changing the context in four transitions, which causes the model to lose the context information of the entity. However, different from the EDA, the MDA utilizes the medical knowledge graph to augment the data, which keeps the structure of sentences unchanged and strengthens the semantic information. Hence, the deep learning models with the MDA can effectively capture the features of sentences to extract entities. Moreover, the BITEmrNER dataset that is compared with the CCKS2019 dataset has complex electronic medical records. However, the results of the F1-score in Table 5 show that the EDA outperforms the original data in entity recognition by 0.88% and the MDA achieves 2.84% improvements in F1-score over all models. Hence, the MDA can alleviate the problem of identifying specialized vocabularies better than EDA.

Table 5. F1-score of the NER task on the CCKS2020 dataset and the BITEmrNER dataset.

Models	CCKS2020			BITEmrNER		
	Original	EDA	MDA	Original	EDA	MDA
BERT + CRF	80.56	81.34	83.29	82.67	83.55	85.51
BiLSTM + CRF	81.00	82.04	83.45	83.35	84.22	85.88
Ra-RC	82.87	83.69	85.64	84.29	84.69	86.64

4.3.2. RC Task Evaluation

To evaluate the effectiveness of MDA in the RC task, two datasets are selected to construct the experiments. In the experiments, the MDA and the EDA are used to generate the augmented data, and then the augmented data are combined with the original data to generate the synthetic data. In addition, three deep learning models that are selected to extract triplets for the RC task are the multi-head attention model [48], the ETL-Span model [49], and the CasRel model [50]. First, the multi-head attention model [48] can simultaneously train the entity extraction module and relation extraction module to recognize multiple relations for each entity. Second, the ETL-Span model [49] is designed as a novel tag schema that can transform the extraction task into tagging task. Third, the CasRel model [50] utilizes a cascade framework that can alleviate the problem of overlapping triplets. Finally, the performance of these models on original data and the synthetic data are evaluated.

For the deep learning models, the RoBERTa model is exploited to capture the features in sentences, the pre-trained weights are the same as the weights of the official release. In addition, the batch size is set to 8, the learning rate is set to 1E-5, and the maximum length of the input sentence is set to 128. Moreover, the stopping mechanism that can stop the training process is also adopted in the experiments.

The experimental process of the RC task is the same as the process of the NER task. Specifically, several samples are selected from original data, and these samples are adopted to generate augmented data on the EDA and the MDA with a 40% replacement probability.

After that, the augmented data are fused with original data to generate synthetic data for experiments. For the CHIP2020 dataset, the MDA optimizes the performance of these models compared to the EDA, and it achieves 2.11% improvements in F1-score over the original data. As can be seen from Table 6, the EDA does not improve the performance of the models in the RC task, because traditional text data augmentation abandons the semantic information that is important to extract the triplets in the RC task. However, the methods of the MDA highlight the importance of label data in the original data, and the method of capturing keywords is exploited to stabilize the semantics of sentences. In addition, the BITEmrRC dataset consists of many electronic medical records, and it contains complex technical terms and overlapping triplets. Hence, the results of the F1-score in Table 6 show that the MDA has excellent performance. Especially, the RoBERTa + CasRel model can alleviate the problems of the complex medical knowledge recognition and overlapping triplets with the data augmentation of the MDA.

Table 6. F1-score of the RC task on the CHIP2020 dataset and the BITEmrRC dataset.

Models	CHIP2020			BITEmrRC		
	Original	EDA	MDA	Original	EDA	MDA
Multi-Head attention	47.10	47.92	49.21	70.27	70.90	72.38
ETL-Span	49.46	60.17	51.38	71.8	72.62	74.33
RoBERTa + CasRel	58.27	59.11	60.18	86.01	86.45	89.37

4.3.3. Ablation Experiments

Based on two NLP tasks of the NER and the RC, the ablation experiments pay attention to the contribution of MKR, WI, WS, WD in the MDA. The Ra-RC model on the BITEmrNER dataset in the NER task and the RoBERTa + CasRel on the BITEmrRC dataset in the RC task are selected as models for the ablation experiments. In addition, the experiments successively remove each of the four methods to obtain the effect of each method. Hence, it can be seen from Table 7 that if the MKR method in the MDA is removed, F1 score significantly reduces in two tasks by 2.16% and 2.36%, respectively. Therefore, the results of F1-score verify that the MKR method can effectively augment the text data by utilizing the medical knowledge graph to replace the keywords. However, after removing the methods of WI, WS, and WD, the F1-scores show that the methods have little influence on the performance of the models. Thus, the improvement of the three methods can be explained by abandoning the semantic information of the keywords. In the end, the results of the F1-score are improved by using the data augmentation of MDA, which indicates that the MDA can significantly improve the performance of the deep learning models in NLP tasks.

Table 7. The results of the ablation experiments.

Methods	NER	RC
	F1	F1
Ours	86.64	89.37
-MKR	84.48	87.01
-WI	85.83	88.49
-WS	86.07	88.74
-WD	86.11	88.71

4.3.4. Case Study

To accurately and directly observe the ability of the MDA, the RC task is selected as the case study, and the medical triplets are recognized from the sentences from the BITEmrRC dataset for exploration. The sentence from the BITEmrRC dataset is shown in Table 8 and its augmented data are shown in Table 9, which contains the text data and the label data. First, from the perspective of text data, the sentence is a piece of descriptive language, which not only contains complex medical nouns but also contains the writing rules of medical orders. Hence, the traditional text data augmentation is adverse to keeping the structure of medical terms. Moreover, the rules of medical orders can be broken. However, with the help of the MDA, the context of all nouns is consistent with the semantics of the original data. Moreover, the MDA also increases the diversity of synthetic sentences. To be specific, the original entities of 'hypesthesia' ('感觉减退'), ' transient numbness of the limb' ('短暂性肢体麻木'), and 'internal carotid artery occlusion' ('颈内动脉闭塞') are changed to the augmentation entities of 'confusion' ('视物模糊'), 'blurred vision' ('头昏'), and 'dizziness' ('神志不清'). In addition, the other types of entities are changed, and different words are exploited in the methods of WI, WS, and WD. Second, from the perspective of label data, the label data changes its corresponding entity pair as the augmentation sentence changes, such as 'Glipizide Sustained Release Capsules' ('唐贝克') and 'Benaglutide Injection' ('贝纳鲁肽'). In addition, the extraction results of the original sentence and the augmented sentence can be obtained from Figures 3 and 4. Especially, the deep learning models accurately identifies nine augmented triplets that includes three diseases, four symptoms, two examination, and three drugs. In the end, the results of the case study fully prove the excellent performance of the MDA in Chinese medical datasets.

Table 8. Original data of the case study on RC task.

Text	Label
脑梗死患者出现左侧肢体麻木, 伴感觉减退, 伴眩晕, 短暂性肢体麻木, 颈内动脉闭塞。遂至我院门诊就诊, 查CT示桥脑梗死可能, 建议MR检查。右侧丘脑陈旧性梗死灶。高血压服用"拜新同, 安博诺"。糖尿病服用"唐贝克"。 (Patients with cerebral infarction developed left limb numbness, hypesthesia, transient numbness of the limb, vertigo, and internal carotid artery occlusion occurred. Then he went to the outpatient department of our hospital. CT examination showed possible pontine infarction, and MR examination was suggested. This is an old infarct in the right thalamus. High blood pressure takes 'Nifedipine Controlled-release Tablets, COAPROVEL'. Diabetes takes 'Glipizide Sustained Release Capsules'.)	脑梗死—临床表现—眩晕, 脑梗死—临床表现—感觉减退, 脑梗死—临床表现—短暂性肢体麻木, 脑梗死—临床表现—颈内动脉闭塞, 脑梗死—影像学检查—CT, 脑梗死—影像学检查—MR检查, 糖尿病—药物治疗—唐贝克, 高血压—药物治疗—拜新同, 高血压—药物治疗—安博诺 ('cerebral infarction'—'clinical manifestation'—'vertigo', 'cerebral infarction'—'clinical manifestation'—'hypesthesia', 'cerebral infarction'—'clinical manifestation'—'transient numbness of the limb', 'cerebral infarction'—'clinical manifestation'—'internal carotid artery occlusion', 'cerebral infarction'—'imaging examination'—'CT', 'cerebral infarction'—'imaging examination'—'MR examination', 'diabetes'—'drug treatment'—'Glipizide Sustained Release Capsules', 'high blood pressure'—'drug treatment'—'Nifedipine Controlled-release Tablets', 'high blood pressure'—'drug treatment'—'COAPROVEL')

Table 9. Augmented data of the case study on RC task.

Text	Label
脑梗死患者产生左侧肢体麻木, 伴视物模糊, 伴严重眩晕, 头昏, 神志不清。遂至我院~~门诊就诊~~, 查CT示桥脑梗死, 建议脑血管造影。梗死灶右侧丘脑陈旧性。高血压直接服用"拜新同, 安博诺"。糖尿病服用"贝纳鲁肽"。 (Patients with cerebral infarction developed left limb numbness, confusion, blurred vision, severe vertigo, and dizziness occurred. Then he went to our hospital. CT examination showed possible pontine infarction, and cerebral angiography was suggested. This is in the right thalamus an old infarct. High blood pressure takes 'Nifedipine Controlled-release Tablets, COAPROVEL'. Diabetes takes 'Benaglutide Injection'.	脑梗死—临床表现—眩晕, 脑梗死—临床表现—视物模糊, 脑梗死—临床表现—头昏, 脑梗死—临床表现—神志不清, 脑梗死—影像学检查—CT, 脑梗死—影像学检查—脑血管造影, 糖尿病—药物治疗—贝纳鲁肽, 高血压—药物治疗—拜新同, 高血压—药物治疗—安博诺 ('cerebral infarction'—'clinical manifestation'—'vertigo', 'cerebral infarction'—'clinical manifestation'—'blurred vision', 'cerebral infarction'—'clinical manifestation'—'dizziness', 'cerebral infarction'—'clinical manifestation'—'confusion', 'cerebral infarction'—'imaging examination'—'CT', 'cerebral infarction'—'imaging examination'—'cerebral angiography', 'diabetes'—'drug treatment'—'Benaglutide Injection', 'high blood pressure'—'drug treatment'—'Nifedipine Controlled-release Tablets', 'high blood pressure'—'drug treatment'—'COAPROVEL')

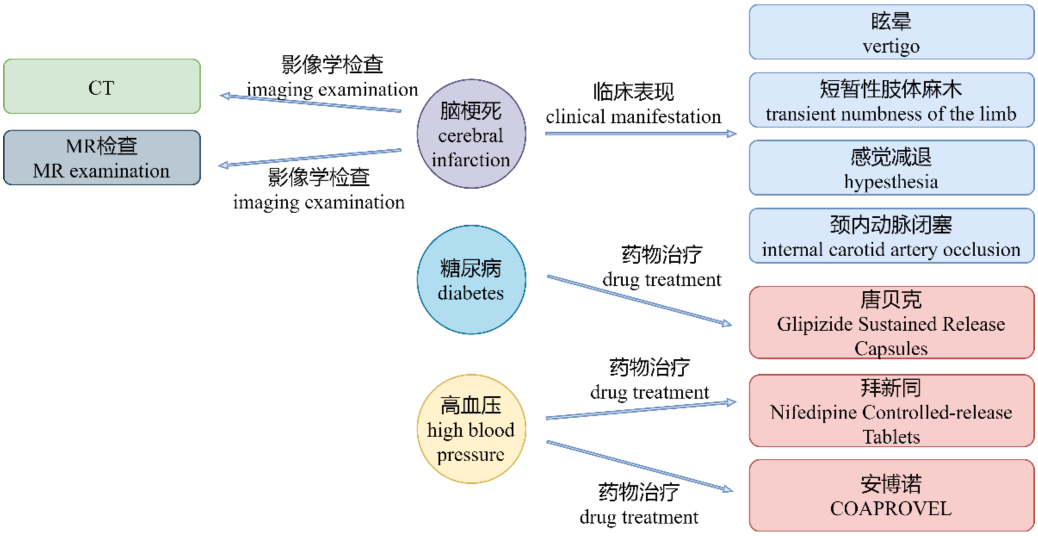

Figure 3. The results of the RC task in original data.

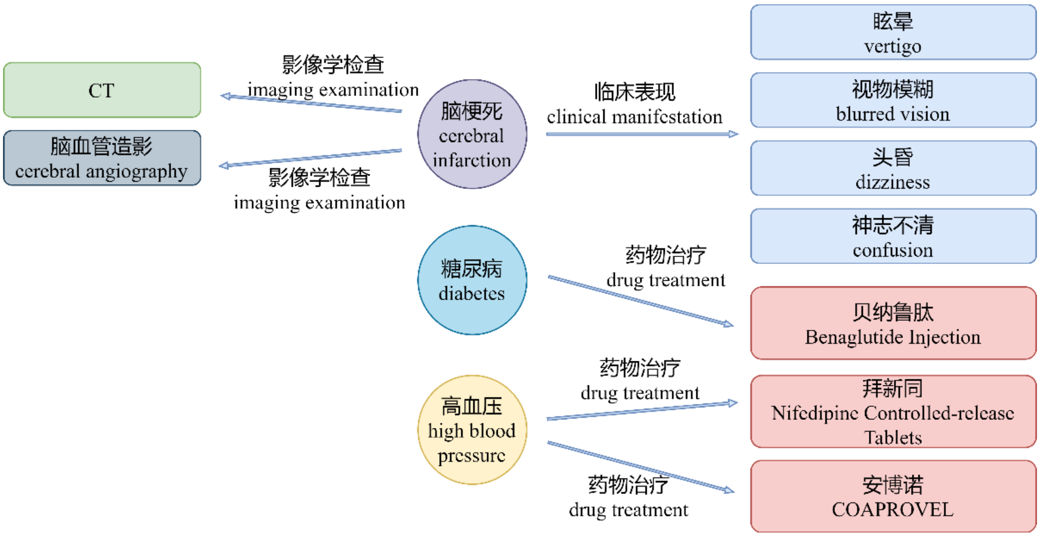

Figure 4. The results of the RC task in augmented data.

4.4. Engineering Applications

The development of intelligent medical treatment cannot be separated from the support of medical data. With the popularization of information technology, more and more hospitals are exploiting electronic medical records in the medical system. Moreover, these electronic medical records not only record the medical data of patients but also make disease predictions for patients. In fact, the application of electronic medical record information in the medical field is a crucial section to promote the sharing of medical data. Because electronic medical records contain different types of data, such as text data and image data, and the text data are mainly exploited in deep learning models to complete the task of NLP. However, deep learning models heavily rely on a large amount of labeled data. In addition, electronic medical records require manual labeling and standardization of standards. In this regard, a medical text data augmentation method that relies on medical knowledge graph is proposed to generate a large number of label data without manual annotation, which is aimed at datasets with a small number of samples. As shown in Figure 5, we can employ a data-driven approach to import some of the electronic medical records data into our methods. Moreover, the MDA method is adopted to generate new synthetic data based on the type of task, which can promote the sharing of medical data.

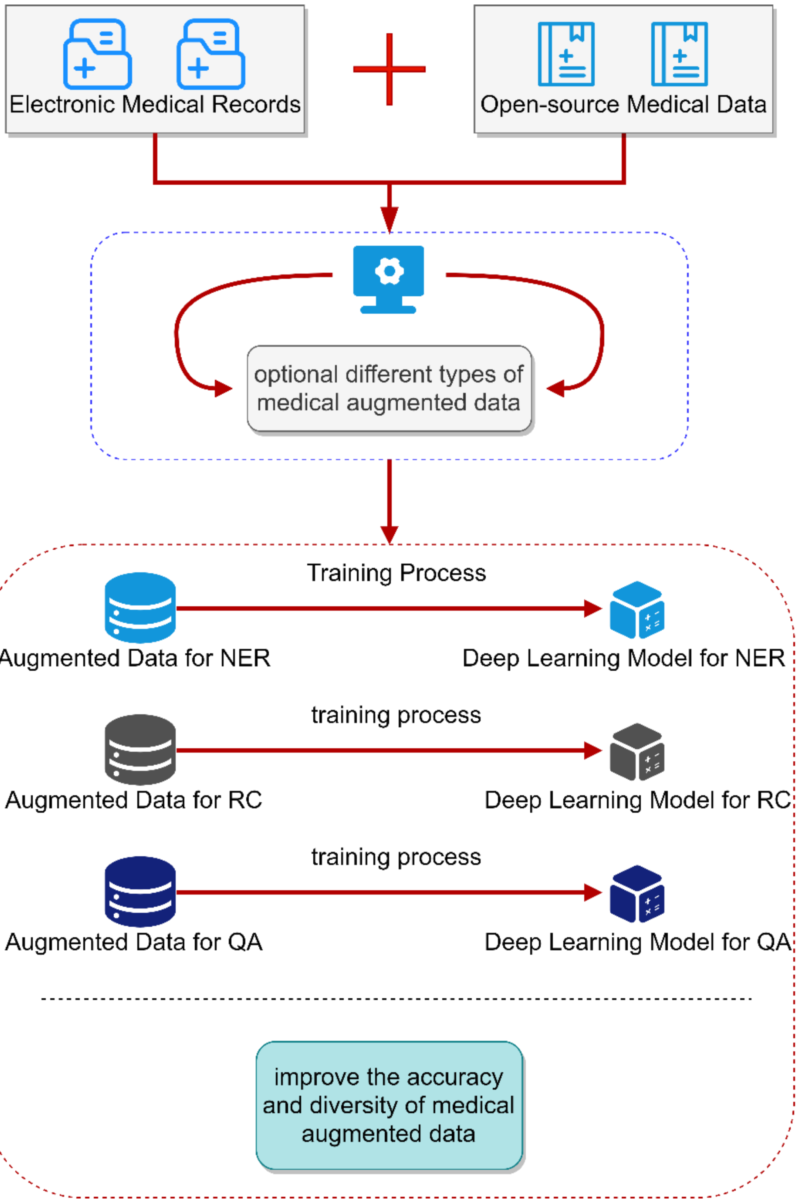
Electronic Medical Records
Open-source Medical Data
optional different types of medical augmented data
Training Process
Augmented Data for NER
Deep Learning Model for NER
training process
Augmented Data for RC
Deep Learning Model for RC
training process
Augmented Data for QA
Deep Learning Model for QA
improve the accuracy and diversity of medical augmented data

Figure 5. An application diagram of the MDA in natural language processing tasks.

As mentioned above, to further enhance the validity of augmented data, our next step is to increase the diversity of sentences by extracting the features of words. Therefore, the language model can be introduced to identify the features of words, which can improve the efficiency of the model for keyword recognition. Furthermore, the MDA can promote the development of the intelligent hospital.

5. Conclusions and Our Future

In this work, we propose a novel medical data augmentation for NLP tasks. The medical knowledge graph based on a large amount of medical knowledge is constructed, which can provide data support for data augmentation. Furthermore, for different downstream tasks, the medical data augmentation can augment the original dataset into augmented dataset that can be applied to the different tasks. In addition, medical data augmentation overcomes the problems of diversity and semantic discontinuity. We conduct complex experiments on four Chinese datasets for two NLP tasks to verify the effectiveness of medical data augmentation. The experiment results for different NLP tasks show that the MDA can be adapted to different tasks, and it outperforms other methods in F1-scores.

Moreover, the ablation experiments demonstrate that each method in MDA can improve the performance of the deep learning models. In summary, the experiments show that the sentences generated by MDA are diverse and can keep the consistency of text data and label data.

In the future, different neural network models will be introduced to extract the features of sentences to identify the keywords more accurately. In addition, these features will be identified to obtain more keywords, and more medical knowledge will be combined to improve the diversity of sentences.

Author Contributions: Conceptualization, B.S. and L.Z. (Lijuan Zhang); methodology, B.S. and H.Z.; software, B.S. and J.W.; validation, B.S. and L.Z. (Lijuan Zhang); formal analysis, B.S.; investigation, B.S.; resources, L.Z. (Lei Zhang) and J.H.; data curation, B.S.; writing—original draft preparation, B.S.; writing—review and editing, L.Z. (Lijuan Zhang); visualization, B.S.; supervision, L.Z. (Lijuan Zhang) and J.H.; project administration, L.Z. (Lei Zhang) and J.H.; funding acquisition, L.Z. (Lei Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: The research was partially funded by Zhejiang Province Key Research and Development Project (2020C03071, 2021C03145).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Huang, W.; Qian, T.; Lyu, C.; Zhang, J.; Jin, G.; Li, Y.; Xu, Y. A multitask learning approach for named entity recognition by exploiting sentence-level semantics globally. *Electronics* **2022**, *11*, 3048. [CrossRef]
2. Hu, W.; He, L.; Ma, H.; Wang, K.; Xiao, J. Kgner: Improving chinese named entity recognition by bert infused with the knowledge graph. *Appl. Sci.* **2022**, *12*, 7702. [CrossRef]
3. Liu, J.W.B.; Su, S. The effect of data augmentation methods on pedestrian object detection. *Electronics* **2022**, *11*, 3185. [CrossRef]
4. Vu, D.T.; Yu, G.; Lee, C.; Kim, J. Text data augmentation for the korean language. *Appl. Sci.* **2022**, *12*, 3425. [CrossRef]
5. Bayer, M.; Kaufhold, M.-A.; Reuter, C. A survey on data augmentation for text classification. *ACM Comput. Surv.* **2022**. [CrossRef]
6. Kumar, P.; Kumar, R.; Srivastava, G.; Gupta, G.P.; Tripathi, R.; Gadekallu, T.R.; Xiong, N.N. Ppsf: A privacy-preserving and secure framework using blockchain-based machine-learning for iot-driven smart cities. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 2326–2341. [CrossRef]
7. Fu, A.; Zhang, X.; Xiong, N.; Gao, Y.; Wang, H.; Zhang, J. Vfl: A verifiable federated learning with privacy-preserving for big data in industrial iot. *IEEE Trans. Ind. Inform.* **2020**, *18*, 3316–3326. [CrossRef]
8. Gao, Y.; Xiang, X.; Xiong, N.; Huang, B.; Lee, H.J.; Alrifai, R.; Jiang, X.; Fang, Z. Human action monitoring for healthcare based on deep learning. *IEEE Access* **2018**, *6*, 52277–52285. [CrossRef]
9. Lejeune, G.; Brixtel, R.; Doucet, A.; Lucas, N. Multilingual event extraction for epidemic detection. *Artif. Intell. Med.* **2015**, *65*, 131–143. [CrossRef]
10. Mounsey, A.; Khan, A.; Sharma, S. Deep and transfer learning approaches for pedestrian identification and classification in autonomous vehicles. *Electronics* **2021**, *10*, 3159. [CrossRef]
11. Asai, M.; Tang, Z. Discrete word embedding for logical natural language understanding. *arXiv* **2020**, arXiv:2008.11649.
12. Funkner, A.A.; Zhurman, D.A.; Kovalchuk, S.V. Extraction of temporal structures for clinical events in unlabeled free-text electronic health records in russian. In *Applying the FAIR Principles to Accelerate Health Research in Europe in the Post COVID-19 Era*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 55–56.
13. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
14. Cheng, H.; Xie, Z.; Shi, Y.; Xiong, N. Multi-step data prediction in wireless sensor networks based on one-dimensional cnn and bidirectional lstm. *IEEE Access* **2019**, *7*, 117883–117896. [CrossRef]
15. Wu, X.; Lv, S.; Zang, L.; Han, J.; Hu, S. Conditional bert contextual augmentation. In Proceedings of the International Conference on Computational Science, Faro, Portugal, 12–14 June 2019; Springer: Berlin/Heidelberg, Germany; pp. 84–95.
16. Li, B.; Hou, Y.; Che, W. Data augmentation approaches in natural language processing: A survey. *AI Open* **2022**, *3*, 71–90. [CrossRef]
17. Daval-Frerot, G.; Weis, Y. Wmd at semeval-2020 tasks 7 and 11: Assessing humor and propaganda using unsupervised data augmentation. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 12–13 December 2020; pp. 1865–1874.

18. Coulombe, C. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv* **2018**, arXiv:1812.04718.
19. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6256–6268.
20. Zhang, Y.; Ge, T.; Sun, X. Parallel data augmentation for formality style transfer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3221–3228.
21. Digamberrao, K.S.; Prasad, R.S. Author identification on literature in different languages: A systematic survey. In Proceedings of the 2018 International Conference on Advances in Communication and Computing Technology (ICACCT), Kochi, India, 13–15 September 2018; IEEE: Piscataway, NJ, USA; pp. 174–181.
22. Perevalov, A.; Both, A. Augmentation-based answer type classification of the smart dataset. In Proceedings of the SMART@ ISWC, Online, 5 November 2020; pp. 1–9.
23. Bornea, M.; Pan, L.; Rosenthal, S.; Florian, R.; Sil, A. Multilingual transfer learning for qa using translation as data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 12583–12591.
24. Wang, W.Y.; Yang, D. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2557–2563.
25. Lowell, D.; Howard, B.; Lipton, Z.C.; Wallace, B.C. Unsupervised data augmentation with naive augmentation and without unlabeled data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 4992–5001.
26. Hou, Y.; Liu, Y.; Che, W.; Liu, T. Sequence-to-sequence data augmentation for dialogue language understanding. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 1234–1245.
27. Li, K.; Chen, C.; Quan, X.; Ling, Q.; Song, Y. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7056–7066.
28. Liu, D.; Gong, Y.; Fu, J.; Yan, Y.; Chen, J.; Lv, J.; Duan, N.; Zhou, M. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space. In Proceedings of the EMNLP (1), Online, 16–20 November 2020.
29. Wu, C.; Ju, B.; Wu, Y.; Lin, X.; Xiong, N.; Xu, G.; Li, H.; Liang, X. Uav autonomous target search based on deep reinforcement learning in complex disaster scene. *IEEE Access* **2019**, *7*, 117227–117245. [CrossRef]
30. Yan, G.; Li, Y.; Zhang, S.; Chen, Z. Data augmentation for deep learning of judgment documents. In Proceedings of the International Conference on Intelligent Science and Big Data Engineering, Nanjing, China, 17–20 October 2019; Springer: Berlin/Heidelberg, Germany; pp. 232–242.
31. Longpre, S.; Wang, Y.; DuBois, C. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4401–4411.
32. Yu, S.; Yang, J.; Liu, D.; Li, R.; Zhang, Y.; Zhao, S. Hierarchical data augmentation and the application in text classification. *IEEE Access* **2019**, *7*, 185476–185485. [CrossRef]
33. Xie, Z.; Wang, S.I.; Li, J.; Lévy, D.; Nie, A.; Jurafsky, D.; Ng, A.Y. Data noising as smoothing in neural network language models. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
34. Guo, H.; Mao, Y.; Zhang, R. Augmenting data with mixup for sentence classification: An empirical study. *arXiv* **2019**, arXiv:1905.08941.
35. Cheng, Y.; Jiang, L.; Macherey, W.; Eisenstein, J. Advaug: Robust adversarial augmentation for neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5961–5970.
36. Min, J.; McCoy, R.T.; Das, D.; Pitler, E.; Linzen, T. Syntactic data augmentation increases robustness to inference heuristics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 2339–2352.
37. Kang, D.; Khot, T.; Sabharwal, A.; Hovy, E. Adventure: Adversarial training for textual entailment with knowledge-guided examples. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2418–2428.
38. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Do not have enough data? Deep learning to the rescue! In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 34, pp. 7383–7390.
39. Quteineh, H.; Samothrakis, S.; Sutcliffe, R. Textual data augmentation for efficient active learning on tiny datasets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 7400–7410.
40. Ng, N.; Cho, K.; Ghassemi, M. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020; Association for Computational Linguistics (ACL): Stroudsburg, PA, USA; pp. 1268–1283.
41. Thakur, N.; Reimers, N.; Daxenberger, J.; Gurevych, I. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 296–310.

42. Yakunin, K.; Mukhamediev, R.I.; Zaitseva, E.; Levashenko, V.; Yelis, M.; Symagulov, A.; Kuchin, Y.; Muhamedijeva, E.; Aubakirov, M.; Gopejenko, V. Mass media as a mirror of the covid-19 pandemic. *Computation* **2021**, *9*, 140. [CrossRef]
43. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; et al. Dbpedia—A large-scale, multilingual knowledge base extracted from wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
44. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196.
45. Huang, Z.; Xu, W.; Yu, K. Bidirectional lstm-crf models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
46. Li, X.; Zhang, H.; Zhou, X. Chinese clinical named entity recognition with variant neural structures based on bert methods. *J. Biomed. Inform.* **2020**, *107*, 103422. [CrossRef] [PubMed]
47. Wu, Y.; Huang, J.; Xu, C.; Zheng, H.; Zhang, L.; Wan, J. Research on named entity recognition of electronic medical records based on roberta and radical-level feature. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 2489754. [CrossRef]
48. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* **2018**, *114*, 34–45. [CrossRef]
49. Yu, B.; Zhang, Z.; Shu, X.; Wang, Y.; Liu, T.; Wang, B.; Li, S. Joint extraction of entities and relations based on a novel decomposition strategy. *arXiv* **2019**, arXiv:1909.04273.
50. Wei, Z.; Su, J.; Wang, Y.; Tian, Y.; Chang, Y. A novel cascade binary tagging framework for relational triple extraction. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 1476–1488.