SNN-BS: A Clinical Terminology Standardization Method Using Siamese Networks with Batch Sampling Strategy

Xiao Wei, Xiaoxin Wang, and Nengjun Zhu(✉)

School of Computer Engineering and Science, Shanghai University, 333 Nanchen Road, Baoshan District, 200444 Shanghai, China
{xwei,wangxiaoxin,zhu_nj}@shu.edu.cn

Abstract. Clinical terminology standardization is important for effective integration and sharing of medical information. It aims to convert clinical colloquial descriptions into standard clinical terminologies. However, the accuracy and efficiency of this task are challenged by the gap between colloquial descriptions and standard terminologies, the slight discrepancy across standard terminologies, and the low efficiency of terminology retrieval. To address these challenges, we propose a novel method called SNN-BS for standardizing clinical terminology based on a Siamese network with a batch sampling strategy. SNN-BS enhances its discrimination ability by sampling a set of terminologies to form a retrieval set with the target terminology. By combing two kinds of similarities, we amplify the differences in features between colloquial descriptions and clinical terminologies while considering deeper semantic relationships. Moreover, we use the lighter Bert-tiny model to encode the terminologies and improve the efficiency of terminology retrieval by reducing comparison numbers through regarding it as a question-and-answer selection task. Finally, we conducted experiments on two datasets to evaluate the performance of our model. The experimental results demonstrate that our method achieves a high level of accuracy, reaching 91.30% and 90.24%, respectively, which outperforms the baselines.

Keywords: Clinical terminology standardization · Siamese network · Batch sampling strategy

1 Introduction

Efficient processing of clinical medical texts using intelligent technologies has become a hot topic in recent years, with standardization of clinical terminology serving as its cornerstone. Its target is to transform the spoken description in clinical medicine (i.e., the origin word) into a standardized description (i.e., the target terminology) in the ICD (International Classification of Diseases) standard terminologies coding set. The standard terminology set classifies diseases according to certain rules based on certain characteristics of clinical diseases

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 272–287, 2023.
https://doi.org/10.1007/978-3-031-46664-9_19

and uses coding methods to represent them. It plays a key role in integrating, exchanging, sharing, and statistics of medical information [14]. Table 1 lists several examples of clinical terminology standardization tasks.

Table 1. Standardized examples of clinical terms

Origin word	Target terminology
HIFU (HIFU for primary liver cancer)	(Ultrasonic scalpel therapy for liver damage)
(Left Ventricular Drainage)	(Ventricular extracranial shunt)
DJ (Pull out DJ tube under cystoscope)	D-J (Cystoscopy D-J tube extraction)

Although clinical terminology standardization has achieved some progress in recent years, existing methods still encounters some challenges, which significantly limit the accuracy and efficiency of this task: 1) The oral expression of the same target terminology is various, and some of them may vary significantly in grammar. It is challenging to associate the origin word solely by analyzing identical tokens without related clinical knowledge. 2) The standard terminologies within the same clinical category often appear very similar, resulting in confusion and difficulty in distinguishing them accurately. The standard terminologies in the same category are mistakenly linked to the same origin word if the origin word is not clearly described. 3) The vast amount of terminologies challenges the efficiency of terminologies standardization. When performing terminology standardization tasks, matching and retrieving target terminologies consume much time, making it challenging to meet the efficiency requirements of clinical medicine.

Existing methods has mainly used the traditional text similarity or semantic similarity evaluation to solve these problems recently [8,13]. However, the degree of specialization in clinical medicine is high, and standard terminologies within the same major category have close similarities. Thus, these methods cannot distinguish similar standard terminologies. Moreover, these methods usually use the form of "$[CLS]o[SEP]d_i[SEP]$" as a standard pair, where the standard terminology $d_i \in D$. Once terminology task requires encoding and comparative learning of $|D|$ standard pairs. While taking the ICD-9 coding set as an example, $|D| \approx 10000$, such methods cannot meet the efficiency requirements of clinical terminology standardization, much less the ICD standard terminology set is constantly expanding.

Our research falls into the category of semantic similarity modeling. To address the mentioned issues, we propose a method named SNN-BS for standardizing clinical terminology based on Siamese networks with batch sampling strategy. The Siamese network has two Siamese subnets with the same structure and shares parameters on the left and right. It inputs two pieces of similar text through the left and right sub-net, and maps them into a new space for feature representation and comparison [7]. It has natural advantages for determining the semantic similarity of the same type of text. However, the traditional Siamese network's approach to this task is not different from other semantic similarity methods. Also, it lacks good recognition ability for relatively similar terminology standard terminologies, which still cannot complete the challenges.

Therefore, SNN-BS includes three customized modules, i.e., Data sampling unit, Bert encoder unit, and Simple feature fusion module. The Data sampling unit is responsible for the sampling generation of the training set. It divides the standard terminology set $|D|$ into k candidate terminology sets W_i and fuses the target terminology t into each W_i. This design reduces the number of encoding and comparison learning from $|D|$ times to k times ($k \ll |D|$). Also, the ability of our model to identify similar standard terminologies can be enhanced. The Bert encoder unit is responsible for tokenizing and encoding words and terminologies. We choose Bert-tiny as the encoder. Its advantage is that the smaller the parameter number of this model is, the faster the inference speed will be. Compared with other methods, it can greatly optimize the efficiency problem and minimize the loss of model accuracy due to its excellent migration effect on the Bert model. The Similar feature fusion module is responsible for calculating the similarity between the origin word o and D_i by combing two kinds of similarities. We use a two-layer similarity calculation method to describe the similarity between clinical texts from different dimensions and through the feed-forward neural network fusion. It can effectively alleviate the problem of low accuracy caused by the large semantic difference between origin word and target terminology. By integrating the three parts of the above design, we can address the above-mentioned challenges and achieve a good balance between the accuracy and efficiency of terminology standardization.

In this paper, our work makes the following contributions: 1) The design of data sampling and Bert encoder unit in SNN-BS also care about the efficiency issue when optimizing the accuracy of clinical terminology standardization. Our method reduces the number of coding and comparison learning by batch sampling the standard terminologies. We adopted a lighter Bert-tiny model with better volume and transfer effect for encoding, which considers the accuracy and efficiency of terminology standardization. 2) We propose a method of randomly mixing the target terminology into each sampling candidate terminology set to enhance the model's ability to select answers for confusing standard terminologies. Besides, we highlight the differences between the origin word and target terminology through two-layer similarity fusion, which can alleviate the problem of low accuracy caused by the highly specialized clinical terminology standardization task and improve the ability to capture long-distance semantic standard terminologies. 3) To better evaluate the method's effect, we tested our model on the Yidu-N7K and the self-built ICD9-INT dataset. The experimental results show that the accuracy of our method in this paper has reached 91.30% and 90.24%, both exceeding the SOTA model in accuracy rate.

2 Related Work

Aiming at the standardization task of clinical terminology standardization, academia mainly adopts two methods based on text similarity and semantic similarity in many academic papers.

Yan et al. [11] introduced a deep generative model to generate the core semantics of the description text and obtained a standard terminology candidate set, and then used the BERT-based semantic similarity algorithm to reorder the candidate set to obtain the final standard terminology. Huang et al. [3] proposed a method for standardizing origin words based on combined semantic similarity technology. It is mainly based on domain knowledge base combined with word segmentation, entity recognition and word vector representation technology to calculate the similarity between origin word and standard terminology. Devlin [1] proposed a pre-trained model BERT, which can predict the similarity between sentence pairs through text classification. Sun et al. [8] proposed to select candidate terminologies based on the Jaccard similarity algorithm, and obtain standard terminology-matching results based on the Bert model. Liu et al. [4] used the method of N-gram and Bert to optimize the solution of many-to-many matching between origin words and target terminologies in ICD-9 encoding.

According to the current results, searching for target terminology of clinical terms by mining semantic information focuses on text structure, but there are knowledge errors in the over-spoken description that are difficult to mine; The similarity comparison method lacks support from clinical expertise, making it challenging to distinguish between easily confused standard terminologies. It can be seen that terminology standardization has been regarded as a natural language processing task, but there is still much room for improvement in terms of semantics or similarity calculation.

Therefore, we propose using the Siamese network that can take into account both semantic information and distinguish the confusing words in the professional field. We will combine this with the Bert-tiny model to complete the encoding of clinical terminology.

3 SNN-BS Method

Our SNN-BS method incorporates two sub-nets (i.e., left-subnet and right-subnet) based on the paradigm of Siamese networks, in which some learnable parameters are shared to encode origin words and standard terminology set, respectively. It solves the problem by adopting a batch sampling strategy and combining Euclidean similarity and dot product similarity to calculate the absolute distance and relative distance between them, which alleviates the problem of significant differences in origin word expression of the same target terminology and high similarity of standard terminologies in the medical field. Besides, we use the Binary Cross entropy loss function to abstract the semantic similarity problem of the clinic terminology standardization task into a question-and-answer selection, i.e., the origin word as a question, and each candidate terminology set as an answer set. This approach allows the model to focus on the detailed differences between each answer while making it easier for parallel operations to improve recognition efficiency.

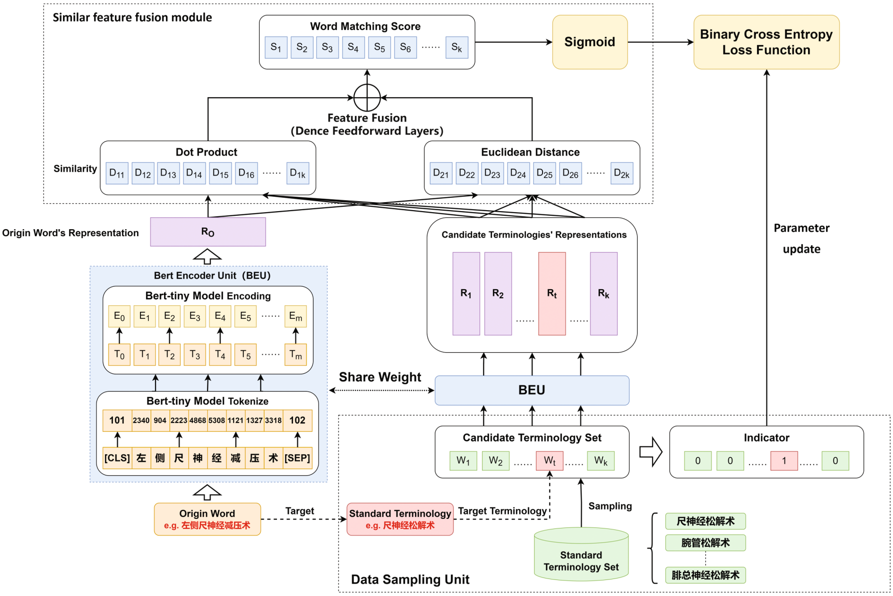
Similar feature fusion module
Word Matching Score
S_1 S_2 S_3 S_4 S_5 S_6 ...... S_k
Sigmoid
Binary Cross Entropy Loss Function
Feature Fusion
(Dence Feedforward Layers)
Similarity
Dot Product
D_{11} D_{12} D_{13} D_{14} D_{15} D_{16} ...... D_{1k}
Euclidean Distance
D_{21} D_{22} D_{23} D_{24} D_{25} D_{26} ...... D_{2k}
Origin Word's Representation
R_O
Candidate Terminologies' Representations
R_1 R_2 ...... R_t ...... R_k
Parameter update
Bert Encoder Unit（BEU)
Bert-tiny Model Encoding
E_0 E_1 E_2 E_3 E_4 E_5 ...... E_m
T_0 T_1 T_2 T_3 T_4 T_5 ...... T_m
Bert-tiny Model Tokenize
101 2340 904 2223 4868 5308 1121 1327 3318 102
[CLS] 左 侧 尺 神 经 减 压 术 [SEP]
Share Weight
BEU
Candidate Terminology Set
W_1 W_2 ...... W_t ...... W_k
Indicator
0 0 ...... 1 0
Origin Word
e.g. 左侧尺神经减压术
Target
Standard Terminology
e.g. 尺神经松解术
Target Terminology
Sampling
Standard Terminology Set
尺神经松解术
腕管松解术
腓总神经松解术
Data Sampling Unit

Fig. 1. Overall framework diagram of SNN-BS

Our model includes four parts shown in Fig. 1: the data sampling unit, the Bert encoder unit (BEU), the similar feature fusion module, and the parameter update unit.

3.1 Data Sampling Unit

The data diversity of the ICD standard terminology set is very limited [10]. Therefore, in the Data Sampling Unit, we sample the standard terminology set D according to a specific size and randomly mix the target terminology into each candidate terminology set W_i. This sampling strategy can enhance the training data and meet the model's generalization ability to the expansion requirements of the future ICD standard terminology set and other similar tasks [12]. Figure 2 shows the strategy for data sampling.

We sample the ICD standard terminology set D according to the size of batch (i.e., S_b, default is 512) parameter set in advance. After that, we divide the standard terminology set into k blocks, where: $k = \lceil |D|/(S_b - 1) \rceil$.

After sampling, we insert the target terminology into each block. Then we get k candidate terminology sets. After that, we generate the corresponding

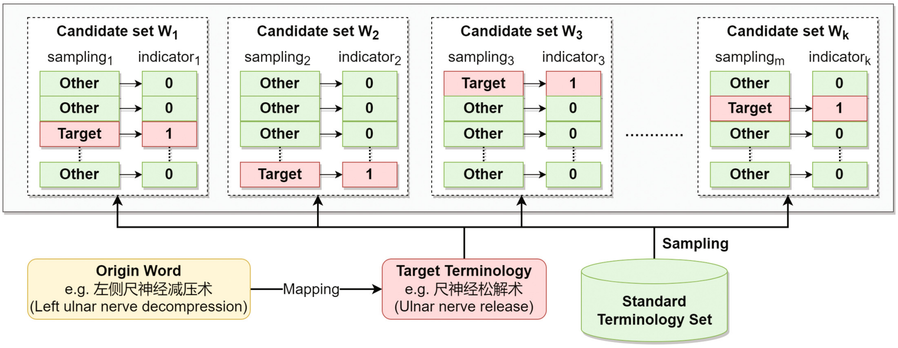

Fig. 2. Sampling strategy for training data: Divide the standard terminology set into multiple candidate terminology set, and randomly insert the target terminology into each set

indicator list as the calculation basis for the subsequent binary cross entropy loss function, where:

$$Indicator_i = \begin{cases} 1 & Candidate_i = TargetTerminology \\ 0 & Candidate_i \neq TargetTerminology \end{cases} \tag{1}$$

We randomly shuffle all the words in each candidate terminology set to ensure the position of the target terminology t in each candidate terminology set has a certain degree of randomness and avoid wrong fitting of the model during loss calculation. Finally we get k candidates word set W_i, where $W_i = D_i \cup \{t\}$. For example, the origin word o input by the user, i.e., “”(“Left ulnar nerve decompression”), we insert its corresponding target terminology t, i.e., “” (“Ulnar nerve release”) at a random position in each candidate terminology set W_i. Then we set the value of t in the corresponding position of the indicator list to 1 and others to 0.

In Bert Encoder Unit (BEU), we tokenize and encode the origin word o and k candidate terminology sets. For example, the origin word “” (“Ulnar nerve release”) will be tokenized through the Bert-tiny model in the form of “$[CLS]$ $[SEP]$”. After obtaining the id of each Chinese character, it is sent to the encoding layer for vectorization processing and input into the Siamese network. During the vectorization process, we set a maximum length limit of 40 characters, and truncation processing will be performed on terminologies longer than the maximum limit.

In the later model training process, the model takes an origin word o and a candidate terminology set W_i as input for a training session. This measure aims to prevent excessive parameter amounts from preventing training when the standard terminology set is too large. Also, it can increase the frequency of occurrence and the comparison of the target terminology, which can effectively strengthen the prediction effect of the model under the limited training set.

3.2 Bert Encoder Unit

The traditional Siamese network structure uses an origin word and a standard terminology as standard pairs for similarity comparison. The input layer's left sub-net and right sub-net are of the same type of text. However, due to the ICD standard terminology set having numerous standard terminologies, using this input structure will result in low recognition efficiency.

Therefore, we redesign the right sub-net input structure and Bert encoder unit of the Siamese network. We take the candidate terminology set as a whole input by the right sub-net and form a standard pair with the origin word o. We use the smaller parameter quantities model "Bert-tiny" as the encoder, which reduces the times of encoding and comparative learning times from the $|D|$ times required by the traditional Siamese network to k times ($k \ll |D|$). This measure can significantly improve the model's efficiency for this task without losing accuracy.

As shown in Fig. 1, we input the origin word o that needs to be standardized into the left sub-net and input the candidate terminology set W_i into the right sub-net. Our method transforms the task of terminology standardization into a question-and-answer selection task by embedding the target terminology t into the input matrix. To encode the input words, we use the Bert-tiny model with a smaller number of parameters to ensure the efficiency of standardization. After encoding the origin word with the Bert Encoder Unit, we vectorize the candidate terminology set W_i input by the right sub-net through the shared parameter weight. In order to avoid the loss of relevance between text and text context, our method adopts the method of Position Embedding to record the position of key information and enhance the dependence between texts.

We use the same weight value and parameters for each set of origin word o and candidate terminology set W_i to tokenize each word according to the Bert-tiny vector representation specification. After this step, each origin word and standard terminology is represented by a S_h-dimensional vector. Where S_h represents the hidden layer size of the pre-training model. We make the following definitions: x represents the vector representation of the origin word, Y_i represents the vector representation of the i-th candidate terminology set, and y_j represents the vector representation of the j-th candidate word in W_i, so we get:$Shape(x) = shape(y_j) = [1, S_h]$, $Shape(Y) = [S_b, S_h]$. For example, the Bert-tiny model's hidden size is 128 dimensions. S_b represents the size of each candidate terminology set.

3.3 Similar Feature Fusion Module

We integrate Euclidean similarity and dot product similarity methods in the Similar Feature Fusion module. When calculating similarity features for vector representations of o and W_i, we also consider the absolute and relative distances between them to further determine the similarity coefficient of the target terminology corresponding to the origin word. Then, we use a fully connected layer to integrate the origin word and target terminology, which highlights the corresponding features between them and improves the capture ability of long-distance semantic target terminologies. Finally, we complete the loss calculation in the training phase.

As shown in Fig. 3, We calculate the absolute and relative distance features between each standard pair. We extract the key information features of them through the dot product similarity and Euclidean similarity so that the model can focus on the key information in the description. Then we calculate the similarity score after the fusion of the two through the fully connected layer.

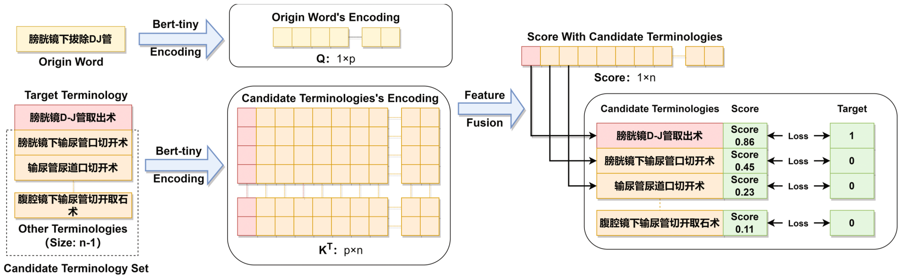

Fig. 3. Feature fusion and loss calculation: The closer the Score value is to 1, the closer the fusion feature of the standard terminology is to the fusion feature of the origin word

Euclidean Distance The Euclidean similarity feature refers to the domain features in clinical terminology descriptors, which can effectively reflect the absolute difference in semantic features between o and each W_i to narrow the selection range of standard terminologies. We calculate the Euclidean distance according to the following formula:

$$Score_{ED}(x, y_j) = \sqrt{\sum_{j+1}^{n}(x - y_j)^2} \tag{2}$$

After that, we get a vector of $[1, S_b]$ dimension with Euclidean similarity feature, where each column in the vector represents the Euclidean distance between the origin word o and the S_b standard terminologies in W_i.

Dot Product The dot product similarity feature measures the individual difference between o and W_i. It can correct the prediction results, which has an enormous description difference between o and W_i. We use its insensitive characteristics to absolute values to filter out the standard terminologies with large differences in distance features with origin word. Let $Q = \{x\}$, $K = Y_i = \{y_1, y_2, ..., y_{512}\}$, then we get: $Shape(Q) = [1, S_h]$, $Shape(K) = [S_b, S_h]$. We calculate the dot product similarity score according to the following formula:

$$Score_{DP}(x, y_i) = softmax(Q \cdot K^T) \tag{3}$$

Similarly, we also get a vector of [1, S_b] dimension with dot product similar feature.

Feature Fusion. To better measure the similarity distance between origin word o and W_i, we set up two layers of full connection layer, align the Euclidean similarity score and the point-product similarity score calculated above, and get the feature vector of [2, S_b] dimension. We use the two-layer feed-forward neural network to map further, express the feature, and output it to the one-dimensional sample space. Finally, we get the similarity score after the vector of origin word x and vector of each candidate word y_i are fused:

$$Matching_i = Contact(Score_{ED}(x, y_i), Score_{DP}(x, y_i)) \tag{4}$$

After obtaining the similar characteristics of the [1, S_b] dimension, we map the matching score of each candidate word to the [0,1] interval through the sigmoid function and record it as $Indicator_j$:

$$Indicator_j = Sigmoid(Matching_i) = \frac{1}{1 + e^{-Matching_j}} \tag{5}$$

For the score of $Indicator_j$, the closer it is to 1, the closer its corresponding candidate word is to the origin word, and vice versa. In Fig. 3, the similar confidence of origin word “” (“Pull out DJ tube under cystoscope”) and candidate word “”(“Cystoscopy D-J tube extraction”) is the highest. Finally, we calculate the loss function and update the weight of the indicator score and indicator target, which is regarded as completing a round of loss calculation.

3.4 Parameter Update Unit

The Parameter update unit is responsible for calculating losses and updating parameters. In the pre-processing stage, we obtain k candidate terminology sets W_i. Each W_i contains target terminology and other interference standard terminologies. Therefore, we need to calculate the indicator score value for each W_i according to the steps in Sect. 3.2. After that, we calculate the loss with the corresponding indicator target value to complete the update of the training parameters.

When selecting a loss function, scholars often use Contrastive loss for loss calculation in the Siamese network. However, in our method, we regard the similarity comparison task as a multi-category task with more categories and use Binary Cross Entropy Loss for optimization learning, which enhances the ability to distinguish similar standard terminologies [6]:

$$BCELoss = -[y \cdot logp(y) + (1 - y) \cdot log(1 - p(y))] \tag{6}$$

where y is the ground truth value corresponding to the origin word, and $p(y)$ is the predicted value of the model output. When $y_i = 1$, the word is the target terminology corresponding to the origin word, then $BCELoss = -logp(1)$. if $p(y_i)$ approaching 1, then the value of $BCELoss$ Approaching 0; if $p(y_i)$ approaching 0, the value of $BCELoss$ Approaching 1, and vice versa. It can be seen that compared with the contrastive loss function commonly used in a Siamese network, binary cross entropy loss has the characteristic of approaching the real label, which plays a key role in the prediction of standard terminologies.

By repeating the above process, one loss calculation for all candidate terminology sets of a single origin word o is regarded as a training batch, and one loss calculation for all origin words is regarded as an epoch.

3.5 Reasoning

The inference structure of our model is similar to the training structure. The standard terminology set is divided into k candidate terminology set W_i. The difference is that our method uses batch sampling and regression prediction to enhance the ability to distinguish confusing words. We take the Top p standard terminologies with the highest confidence for the prediction results of each candidate terminology set W_i to form a new candidate terminology set W_{k+1}, and use it as the input for the next round. Which defines: $p = \lceil S_b/50 \rceil$

For the $k*p$ candidate standard terminologies of k candidate terminology sets, we input them again into the model structure for secondary reasoning to obtain a new matching score. Then the model will output p standard terminologies with the highest confidence. We merge these outputs according to the output number requirements to form the final prediction result.

4 Experimental Results and Analysis

4.1 Dataset

Our test dataset consists of two parts: the YiduN7K dataset from CHIP2019 and the self-built dataset based on the ICD-9 international dictionary set, which is named ICD9-INT.

YiduN7K Dataset. The YiduN7K dataset is one of the clinical medical information processing evaluation tasks from CHIP2019 (China Conference on Health Information Processing). The origin words are all from the real medical data of Grade III A hospitals, including 4000 training word pairs, 1000 validation word pairs, and 2000 testing word pairs. The data structure is also presented in the form of <origin word, target terminology>.

Since there are many-to-many standard terminology prediction entries in the CHIP2019 dataset, the accuracy rate is defined as the total number of origin words to be predicted divided by the combination of origin words and target terminologies:

$$A = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{max(|P_i|, |G_i|)} \tag{7}$$

where P_i is the standard terminology set predicted by the origin word i, and G_i is the real target terminology set of the origin word i.

ICD9-INT Dataset. ICD9-INT dataset is built according to the ICD-9-CM-3 international version coding standard [2]. The ICD-9-CM-3 international version contains 4875 standard terminologies. For each origin word, we use the NLPCDA data enhancement tool to generate two corresponding confusion words, build a 9,750-size dataset, and generate 7,800 training word pairs and 1950 testing word pairs through random segmentation. The data structure is also presented in the form of <origin word, target terminology> pair as shown in Table 1. We still use the formula 7 to calculate the accuracy.

4.2 Main Result

We compare our model with the baseline model of existing methods for ICD terminology standardization tasks:

HCAN [5] completes semantic matching between origin word and target terminology through multi-granularity importance weight measurement and models short text similarity to select target terminology. The ABTSBM model [4] utilizes neural networks to train the original term combination splitting method based on named entity recognition and part of speech tagging for the ICD dataset of many-to-many. The Bert-target method [1] pre-trains bidirectional representations through left and right contexts and selects target terminology through question answering and inference after fine-tuning. The Bert with Longest common sub-sequence method [11] analogizes the clinical terminology standardization task to a translation task. It introduces a deep generative model to generate the core semantics of the description text and reorder the candidate set by using the Bert-based semantic similarity algorithm to obtain the final target terminologies. The Bert with Jaccard algorithm method [8] calculates the Jaccard similarity coefficient between the origin word and target terminology to be standardized. It generates a set of candidate standard terminologies, and uses the Bert model for matching and classification.

To optimize the training process, we set the S_b to 512, the learning rate to 1e-5, and the epochs to 50. The model parameters are saved using the early stop method. For other methods, all parameters are tuned to achieve their best performance. Then, the comparison results are shown in Table 2.

We can have the following main observations:

First, our method outperforms all the compared methods on both datasets. On CHIP2019, it achieves very high accuracy, i.e., 91.30%, which increases that of HCAN by 17.8%. The reason behind this is that compared with simple semantic modeling, using pretrained models can better explore the correlation between the

Table 2. Encoding Efficiency Comparison of Different Pre-training Models

Paper	Method	$Acc_{YiduN7K}$	$Acc_{ICD9-INT}$
Rao et al., 2018 [9]	HCAN	73.50%	74.51%
Devlin et al., 2018 [1]	Bert-target	88.00%	86.10%
Yijia Liu, 2021 [4]	ABTSBM	87.50%	86.61%
YAN Jinghui, 2021 [11]	Longest common sub-sequence	89.00%	87.18%
SUN Yuejun, 2021 [8]	Jaccard algorithm	90.04%	88.82%
Ours	**SNN-BS**	**91.30%**	**90.24%**

origin word and the target terminology. Second, Devlin et al.'s Bert-target model achieved good results at the time, i.e., 88.00%. It utilizes the Bert pre-training model to transform terminology standardization tasks into text classification tasks, and introduces the pre-training model into terminology annotation tasks.

Third, the Bert with Jaccard algorithm has also achieved good results, combining Jaccard algorithm and Bert encoder, which can better utilize the good feature extraction characteristics of Bert encoder in terminology standardization tasks.

Although our model uses the Bert-Tiny model with fewer parameters and lighter weight, it still has improved prediction accuracy compared with the SOTA model. This is mainly due to our redesign of the Siamese network structure for the clinical terminology standardization task, so that by generating multiple loss calculations under the data sample unit, the model is able to fit better to the data. After encoding the origin word and candidate terminology set, we fused the absolute and relative distance features. It makes our model not only consider the explicit similarity of the text, but also capture the deeper semantic features in the clinical terminology vocabulary, which has better generalization ability for the test samples with obvious features but difficult similarity matching.

4.3 Contrast and Ablation Experiment

In order to verify the reasoning ability of our model, we compared Bert-tiny with the common Bert-base and Robert-small model reasoning speed under the same framework. We simulated the efficiency of one-to-one similarity comparison in the SOTA method. Table 3 shows the experimental results.

Table 3. Comparison of Encoding efficiency of different pre-training models

Model	S_b	$Size_{code}$	Time Cost
SNN-BS(+Bert-Tiny)	512	2000	144.42 s
SNN-BS(+Bert-Tiny)	1	2000	5324.52 s
SNN-BS(+Robert-Small)	512	2000	722.79 s
SNN-BS(+Bert-Base)	512	2000	1278.70 s

According to the results in Table 3, it can be seen that using the Bert-tiny model can significantly reduce the efficiency of model reasoning compared with Robert-small and Bert-base. With the same Bert-tiny model, using candidate terminology set as the right-subnet input (set the S_b to 512) has a significant improvement in efficiency compared with using single standard terminology as the right-subnet input (set the S_b to 1).

We conducted ablation experiments for the data sampling unit and similar feature fusion module we adopted. Table 4 shows the experimental results.

Table 4. Comparison of Encoding efficiency of different pre-training models

Model	$Acc_{YiduN7K}$	$Acc_{ICD9-INT}$
SNN-BS(+Dot Product)	84.55%	84.10%
SNN-BS(+Euclidean Distance)	86.25%	85.84%
SNN-BS(+Dot Product & Euclidean Distance)	89.50%	89.02%
SNN-BS(+Data Sampling & Dot Product)	86.65%	86.21%
SNN-BS(+Data Sampling & Euclidean Distance)	88.05%	87.48%
SNN-BS(+All)	**91.30%**	**90.24%**

According to the results in Table 4, it can be seen that the accuracy of the method using the data sampling unit has increased by about 1.78% compared with the method without this strategy; The accuracy rate of only considering dot product similarity is about 3.11% higher than that of considering fusion similarity; The accuracy rate of only considering the Euclidean distance similarity is increased by about 4.64% compared with that of considering the fusion similarity. Furthermore, the model's prediction accuracy can reach the best effect of 91.30% when using the data sampling unit and similar feature fusion module.

4.4 Case Study

For the target terminology results predicted by our method, we selected some origin word samples with the incorrect prediction of SOTA model. We compared the prediction results of Bert-target [1] method, Siamese network without batch sampling strategy, and Siamese network with batch sampling strategy. Table 5 shows the experimental result.

From the result, our method can effectively match those standard pairs with similar meanings between origin words and target terminologies. After the data sampling unit and similar feature fusion module, the ability to distinguish "all" or "local" operations in target terminology can be effectively improved. For origin word with vague body parts, it can also match the corresponding target terminology. For example, the word ""("Right lower leg amputation") should focus on the characteristics of ""("thigh") rather than ""("lower body"). However, in some cases where there are ambiguous and mixed descriptions in the

Table 5. Sampling comparison of prediction results

Origin word/Target terminology	Method	Forecast Word	Result
(O)[1]	Bert-target	[7]	False
(T)[2]	SNN-BS(initial)	[2]	True
	SNN-BS(tuning)	[2]	True
(O)[3]	Bert-target	[8]	False
(T)[4]	SNN-BS(initial)	[8]	False
	SNN-BS(tuning)	[4]	True
(O)[5]	Bert-target	[9]	False
(T)[6]	SNN-BS(initial)	[9]	False
	SNN-BS(tuning)	[9]	False

[1]: Endoscopic Assisted Thyroglossal Duct Cyst Resection
[2]: Thyroglossal duct lesion resection
[3]: Right lower leg amputation
[4]: Thigh amputation
[5]: D-J tube implantation under cystoscope
[6]: Transurethral ureteral stenting
[7]: Thyroglossal duct resection
[8]: Lower extremity amputation
[9]: Cystoscopy D-J tube extraction

origin word, its terminology standardization ability still needs to be improved due to lacking external domain knowledge. For example, the origin word, i.e., “(“D-J tube placement under cystoscopy”) is too concerned about the semantic characteristics of “”(“D-J tube”), which leads to the neglect of global semantics in the prediction process, resulting in the wrong result.

5 Conclusion

Aiming to address the accuracy and efficiency problems of clinical terminology standardization task, we reduced the number of encoding and comparison learning by sampling the standard terminology set. We used the Bert-tiny model, which is lighter and has a better migration effect for encoding. We randomly mixed the target terminology into each sampling candidate terminology set to strengthen the model's ability to select answers for confusing standard terminologies. Through two-level similarity fusion, it highlight the corresponding characteristics between the origin word and the target terminology, alleviating the low accuracy problem caused by the strong professionalism of clinical terminology standardization tasks and improving the capture ability of remote semantic standard terminologies.

We take into account the accuracy and efficiency of terminology standardization. The experimental results of this method on the Yidu-N7K dataset and the ICD9-INT dataset show that our method is superior to the SOTA model in accuracy. Also, it can effectively improve the standardization accuracy of the origin word with unclear descriptions of surgical sites. However, the accuracy needs to be improved for test cases with clinical abbreviations in the origin word. In the next step. We can consider combining abbreviations in the clinical medical field and external knowledge bases of abbreviations to further improve the terminology standardization ability of the model.

Acknowledgemment. This work is partially supported by National Natural Science Foundation of China under Grant No. 62202282 and Shanghai Youth Science and Technology Talents Sailing Program under Grant No. 22YF1413700.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
2. Gao, Y., Fu, X., Liu, X., Wu, J.: Multi-features-based automatic clinical coding for Chinese ICD-9-CM-3. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) ICANN 2021. LNCS, vol. 12895, pp. 473–486. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86383-8_38
3. Huang, J.: Automatic encoding of disease terminology based on combined semantic similarity calculation. Micro Comput. Appl. **36**(08), 157–160 (2020)
4. Liu, Y., Li, S., Yu, J., Tan, Y., Ma, J., Wu, Q.: Many-to-many Chinese ICD-9 terminology standardization based on neural networks. In: Huang, D.-S., Jo, K.-H., Li, J., Gribova, V., Hussain, A. (eds.) ICIC 2021. LNCS, vol. 12837, pp. 430–441. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-84529-2_36
5. Rao, J., Liu, L., Tay, Y., Yang, W., Shi, P., Lin, J.: Bridging the gap between relevance matching and semantic matching for short text similarity modeling. 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), pp. 5370–5381 (2019)
6. Ruby, U., Yendapalli, V.: Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends. Comput. Sci. Eng. **9**(10), 1–8 (2020)
7. Sinha, R., Desai, U., Tamilselvam, S., Mani, S.: Evaluation of Siamese networks for semantic code search. arXiv preprint arXiv:2011.01043 (2020)
8. Sun, Y., Liu, Z., Yang, Z., Lin, H.: Standardization of clinical terminology based on BERT. Chinese J. Inf. Technol. **35**(4), 75–82 (2021)
9. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
10. Wullschleger, P., Lionetti, S., Daly, D., Volpe, F., Caro, G.: Auto-regressive self-attention models for diagnosis prediction on electronic health records. In: 2022 IEEE International Conference on Big Data, pp. 1950–1956. IEEE (2022)
11. Yan, J., Xiang, L., Zhou, Y., Sun, J., Chen, S., Xue, C.: Application of deep generative model in clinical terminology standardization. Chinese J. Inf. Technol. **35**(5), 77–85 (2021)
12. Zhang, Z., Liu, J., Razavian, N.: Bert-xml: Large scale automated ICD coding using BERT pretraining. arXiv preprint arXiv:2006.03685 (2020)

13. Zhou, L., Qu, W., Wei, T., Zhou, J., Gu, Y., Li, B.: A review on named entity recognition in Chinese medical text. In: Sun, X., Zhang, X., Xia, Z., Bertino, E. (eds.) ICAIS 2021. CCIS, vol. 1422, pp. 39–51. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78615-1_4

14. Zhu, N., Cao, J., Shen, K., Chen, X., Zhu, S.: A decision support system with intelligent recommendation for multi-disciplinary medical treatment. ACM Trans. Multim. Comput. Commun. Appl. **16**(1s), 33:1-33:23 (2020)