chinaXiv:202010.00060v2

An Advanced ICD-9 Terminology Standardization Method Based on BERT and Text Similarity

Yijia Liu[1,*], Bin Ji[1], Jie Yu[1], Yusong Tan[1], Jun Ma[1] and Qingbo Wu[1]

[1] College of Computer, National University of Defense Technology, Hunan, China
[*] lcf2777655073@163.com

Abstract. The ICD-9 terminology standardization task aims to standardize the colloquial terminology recorded by doctors in medical records into the standard terminology defined in the ninth version of International Classification of Diseases (ICD-9). In this paper, we first propose a **B**ERT and **T**ext **S**imilarity **B**ased **M**ethod (BTSBM) that combines BERT classification model with text similarity calculation algorithm: 1) use the N-gram algorithm to generate a **C**andidate **S**tandard **T**erminology **S**et (CSTS) for each colloquial terminology, which is used as the training dataset and test dataset for next step; 2) use the BERT classification model to classify the correct standard terminology. In this BTSBM method, if a larger-scale CSTS is taken as the test dataset, the training dataset also needs to maintain larger-scale. However, there is only one positive sample in each CSTS. Hence, expanding the scale will cause a serious imbalance in the ratio of positive and negative samples, which will significantly degrade system performance. While if we keep the test dataset relatively small, the CSTS **A**ccuracy (CSTSA) will degrade significantly, which results a very low system performance ceiling. In order to address above problems, we then propose an optimized terminology standardization method, called as **A**dvanced **B**ERT and **T**ext **S**imilarity **B**ased **M**ethod (ABTSBM), which 1) uses a large-scale initial CSTS to maintain a high CSTSA to ensure a high system performance ceiling, 2) denoises CSTS based on body structure to alleviate the imbalance of positive and negative samples without reducing the CSTSA, and 3) introduces the focal loss function to further promote a balance of positive and negative samples. Experiments show that, the precision of the ABTSBM method is up to 83.5%, which is 0.6% higher than BTSBM, while the computation cost of ABTSBM is 26.7% lower than BTSBM.

Keywords: ICD-9 Terminology Standardization, Text Similarity, BERT.

1 Introduction

The International Classification of Diseases (ICD) is an internationally unified disease classification method formulated by the WHO. ICD-9 is its ninth edition, which classifies diseases based on surgical operations. The key components of ICD-9 terminology are body structures and surgical names. "腰椎间盘切除术" ("Lumbar Discectomy") and "关节活组织检查" ("Biopsy of Joint") are two typical instances.

chinaXiv:202010.00060v2

At present, medical terminologies recorded by doctors in medical records often contain information such as abbreviations and colloquialisms. Doctors may also record medical terminologies using excessive fine-grained or coarse-grained descriptions. Simultaneously, when using medical terminologies, hospitals and institutions may use their self-defined standard terminologies. As a result, it sets a heavy barrier for academic communication in medical research field. So it is a real need for hospitals and doctors to map these medical terminologies to unified standard ones. And for medical insurance, a unified name for the same disease recorded in different descriptions benefits to quantify insurance compensation.

In hospitals, manually terminology standardization needs professional knowledge. Due to the huge quantity of ICD-9 standard terminologies, and numerous colloquial terminologies produced by the hospital every day, it is a time-consuming and laborious job. Therefore, the automatic standardization of medical terminology has become an urgent need for hospitals and doctors.

The aim of this paper is to find the corresponding standard ICD-9 terminology for the given original medical terminology (namely the colloquial terminology). We first propose a **B**ERT and **T**ext **S**imilarity **B**ased **M**ethod (BTSBM) that combines the BERT classification model with the text similarity calculation algorithm: 1) Use the N-gram algorithm to filter out ICD-9 standard terminologies that are highly similar to the original terminology. Standard terminologies with top N of the highest similarities (Top-N) are taken to form the **C**andidate **S**tandard **T**erminology **S**ets (CSTSs) for original terminologies, which are taken as training and test dataset for next step; 2) Use the BERT classification model to predict the standard terminology of the original terminology. The reason for not predicting the correct standard terminology with all ICD-9 standard terminologies are that there are too many irrelevant items and the computation cost is too large. Fig.1 shows the framework of BTSBM. Through the BERT classification model, we obtain the BERT predicted negative or positive (i.e. 0 or 1) label of each candidate terminology, and choose the one with the highest probability among all candidate terminologies as the standard terminology.

However, in this BTSBM method, there is an imbalance in the proportion of positive and negative samples. The scale of CSTS is positively correlated with the CSTS **A**ccuracy (CSTSA, which is defined in formula 5), and CSTSA determines the upper limit of system performance. If a larger-scale CSTS is used in the test dataset, in order to have better system performance and robust generalization ability, the training dataset also needs to use a large-scale CSTS. For the standard terminology corresponding to each original terminology is unique, the larger N, the larger negative samples contained in CSTSs, which will cause a serious imbalance in the proportion of positive and negative samples. As a consequence, system performance significantly degrades. While if a small-scale CSTS is used as the test dataset, the low CTSTA sets a low system ceiling performance, which is insupportable both in academic research and real-world application.

In order to address above problems, we propose an optimized terminology standardization method, called as **A**dvanced **B**ERT and **T**ext **S**imilarity **B**ased **M**ethod (ABTSBM): 1) Use large-scale initial CTSTs to maintain a high CSTSA to ensure high system performance ceiling; 2) Use body structure based data denoising technique to

reduce the imbalance and further reduce the computation cost without affecting CSTSA; 3) Use the focal loss function to further solve the imbalance problem in the training dataset, and improve system performance. In result, we efficient alleviate the serious imbalance between positive and negative samples caused by large-scale CSTS.

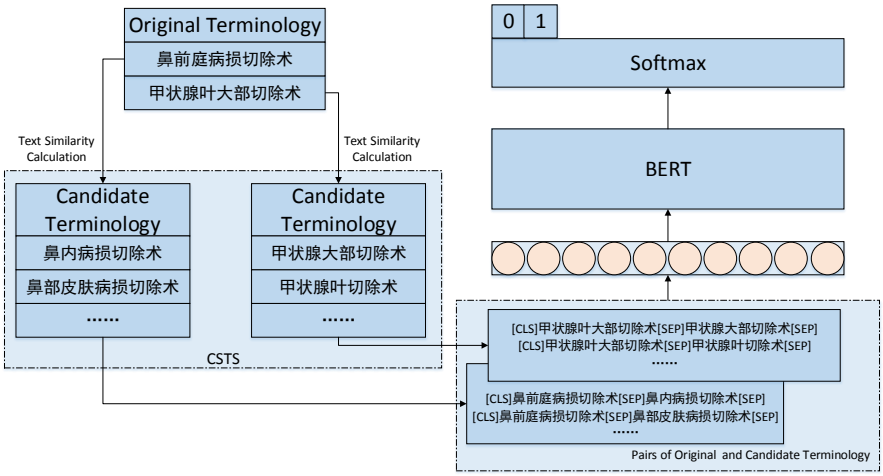

Fig. 1. The framework of BTSBM. Based on the text similarity calculation, the CSTS corresponding to the original terminology is generated. Then, original terminology and its candidate terminology pairs are input into the BERT classification model, and the candidate terminology with the highest probability is output.

2 Related work

Regarding the standardization of ICD Terminology, Liu [1] once developed a complete entry system on ICD10. By standardizing the filling content of doctors, enter standard terminologies directly. Cheng [2] improved a dictionary of the work presented in [1]. However, these methods not only rely on the input of doctors, but also cannot solve the problem of old medical records standardization. The large amount of information contained in old medical records is exactly what doctors cannot ignore when doing research.

We believe that the ICD-9 terminology standardization task can be formalized as a text similarity task based on deep learning. At present, the related work of text similarity includes text similarity calculation algorithm based on string and some methods based on neural network. There are summaries of the methods of text similarity calculation based on string [3-5], such as N-gram [6], Longest Common Subsequence [7] and Edit Distance [8]. Yu et al. [9] used Jaccard Distance to calculate the similarity between two texts. Sidorov et al. [10] proposed an algorithm for tree Edit Distance. The methods based on neural network mainly calculate similarity by generating word vectors. Kenter et al. [11] used word vectors of different dimensions to train the classifier to predict the

chinaXiv:202010.00060v2

similarity score between short texts. Mikolov et al. [12] and Pennington et al. [13] proposed Word2Vec and GloVe to generate word vectors, respectively. Devlin et al. [14] proposed a pre-training model BERT, which predicts the similarity between sentence pairs by 0-1 binary classification of sentence pairs.

3 Method

3.1 BTSBM

As shown in Fig.1, BTSBM composes of two parts: the text similarity and the BERT. The first part uses a string-based text similarity calculation algorithm to get the similarity of ICD-9 standard terminologies and the original terminology. Then, take the ICD-9 standard terminologies with the Top-N highest similarity as a CSTS for each original terminology. The second part uses the BERT classification model to predict the similarity between each candidate terminology and the original terminology, and output the candidate terminology with the highest predicted probability.

In the first part, through the construction of CSTS, we can effectively screen out the terminology in the ICD-9 standard terminologies that is highly similar to the original terminology. It avoids the huge computation cost for the BERT classification model caused by the large number of pairs of each original terminology with the ICD-9 standard terminologies. And it also filters out some interference items that may affect the BERT prediction result.

Our commonly used text similarity calculation algorithms contain N-gram algorithm, Longest Common Subsequence algorithm and Edit Distance algorithm. The basic idea of the N-gram algorithm is to divide the terminology into sub-sequences according to length N, and these sub-sequences are called grams. Then, count the number of the same gram in two strings to measure the similarity.

In BTSBM, we select the N-gram algorithm to calculation the similarity of the original terminology between ICD-9 standard terminologies. Then we can screen the Top-N similarity ICD-9 standard terminologies for constructing the CSTS. The formula of N-gram algorithm is as shown in Formula 1.

$$sim\left(Terminology_i, Terminology_j\right) = \frac{2*Ngram(Terminology_i, Terminology_j)}{len(Terminology_i)+len(Terminology_j)} \quad (1)$$

In the second part, the BERT classification model is in binary classification mode, which is used to predict whether the original terminology is similar to its candidate terminology, as shown in Fig.1. However, because the standard terminology corresponding to an original terminology is unique, so we do not obtain the label outputted by BERT, but to obtain the candidate terminology with the highest probabilities among the probability of candidate terminologies. Two examples are shown in Table 1, in the case of "异体肾移植术" ("Allogeneic Kidney Transplantation"), although there are multiple candidate terminologies predicted as positive and their probabilities were very similar, "肾异体移植术" ("Kidney Allograft Transplantation ") with the highest probability is the correct standard terminology. Also, in the case of "骨盆外固定架固定术"

chinaXiv:202010.00060v2

("Pelvic External Fixation "), even if BERT judges these two terminologies are not similar because of the probabilities less than 0.5, the candidate terminology "盆骨外固定术" ("Pelvic External Fixation") with the highest probability is still chosen as the correct standard terminology.

Table 1. Instances of BERT prediction results.

Original Terminology	Candidate Terminology	Probability of Positive Sample
异体肾移植术	肾异体移植术	0.927
	肾移植	0.922
骨盆外固定架固定术	盆骨外固定术	0.091
	骨盆外固定装置去除术	0.066

3.2 ABTSBM

In BTSBM, if a larger-scale CSTS is used for the test dataset, a larger-scale CSTS is also required for the training dataset to learn enough features to distinguish interference items in test dataset. But there is only one positive sample in CSTS, because each original terminology has only one corresponding standard terminology. And with the expansion of CSTS, the number of negative samples gradually increases, which will cause a serious imbalance in the proportion of positive and negative samples in the training dataset. And it will significantly reduce the system performance. Taking the CSTS constructed by Top30 as an example, the proportion of positive and negative samples is 1:29. Because the precision of the BERT is not equal to the precision of the terminology standardization task, and it is evaluated based on whether the 0-1 label is correct, rather than whether the candidate terminology with the highest probability is the correct standard terminology. So in the extreme case all classified as label 0, the precision of the BERT classification can still reach 97%, and the precision of the terminology standardization task is 0%. Therefore, due to the sparse positive examples, it is difficult to train. However, if the test dataset uses a small-scale CSTS, it will directly affect the CSTSA, and the lower CSTSA determines the upper limit of the system performance, which obviously cannot achieve high system performance.

In order to solve the above problems, we propose ABTSBM, an optimized terminology standardization method, as shown in Fig.2. Compared with BTSBM, we use two methods to alleviate the imbalance of positive and negative sample ratio caused by large-scale dataset: 1) denoise the dataset based on body structure to delete irrelevant candidate terminologies; 2) Use the focal loss function to enhance the BERT classification model’s learning ability for unbalanced training dataset.

chinaXiv:202010.00060v2

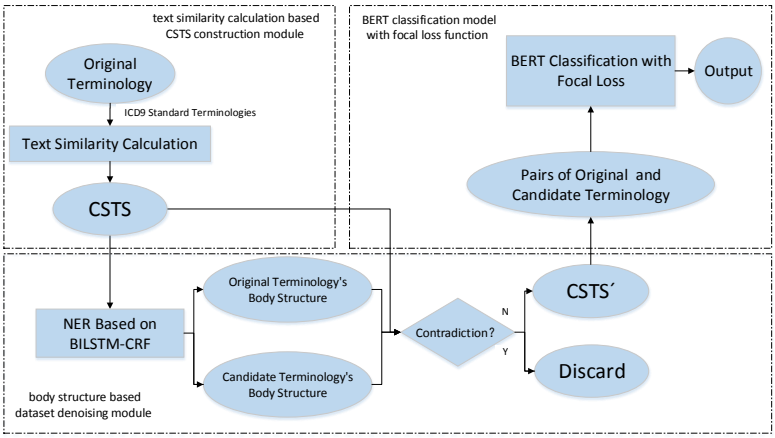

Fig. 2. The framework of ABTSBM, which composes of three parts: text similarity calculation based CSTS construction module; body structure based dataset denoising module; BERT classification model with focal loss function.

Denoising methods. In the CSTS obtained by the N-gram algorithm, there are some candidate terminologies that are highly similar to the original terminologies but are obviously incorrect. Some examples are show in Table 2: in the original terminology "支撑喉镜下声带病损摘除术" ("Extraction of Vocal Cord Lesions under Support Laryngoscope"), the supplementary information "支撑喉镜"("support laryngoscope") does not match the information "内镜"("endoscope") in its standard terminology. But by the N-gram algorithm, "支撑喉镜下声带注射术" ("Vocal Cord Injection under Support Laryngoscope") is more similar than the standard terminology; in the case of "内镜下右侧甲状腺叶切除术" ("Endoscopic Right Thyroid Lobectomy"), the body structure of the candidate terminology is "腺样体" ("Adenoids"), which is obviously contradictory to the body structure "甲状腺叶" ("Thyroid Lobe") of the original terminology. Based on the above data characteristics, we investigate two denoising methods, which are Denoising method based on Supplementary Information (namely DSI) and Denoising method based on Body Structure (namely DBS).

Table 2. Instances of original terminologies and their standard terminologies and candidate terminologies.

Original Terminology	Standard Terminology	Candidate Terminology.
支撑喉镜下声带病损摘除术	内镜下声带病损切除术	支撑喉镜下声带注射术
内镜下右侧甲状腺叶切除术	单侧甲状腺叶切除术	内镜下腺样体切除术

The DSI aims to remove the supplementary information in the original terminology and the ICD-9 standard terminologies (such as approach method, endoscope name and

chinaXiv:202010.00060v2

location word, etc.). As a result, the treated candidate terminology composes of components that can directly determine the standard terminology (Such as body structure and surgical name).

The DBS first uses the BILSTM-CRF model [15] to perform named entity recognition, and to extract the body structures of the original terminology and candidate terminologies. Then, compare the body structures existing in the original terminology and its candidate terminologies, if a candidate terminology contains body structure that is not contained in the original terminology, then the candidate terminology and the original terminology contradict in aspect of body structure, so we discard the candidate terminology. Some examples are given in Table 3, the original terminology "胃穿孔修补术" ("Gastric Perforation Repair") includes the body structure "胃" ("stomach"), and its candidate terminology "肠穿孔修补术" ("Bowel Perforation Repair") includes the body structure "肠"(" bowel "). Because of body structure contradiction, this candidate terminology is discarded. And the other candidate terminology "胃修补术" ("Gastric Repair") contains the same body structure as the original terminology, so it is retained. We manually annotate ICD-9 standard terminologies with BIO tagging scheme to obtain training dataset for BILSTM-CRF.

Table 3. The Instances of the denoising method based on body structure.

Original Terminology	Body Structure	Candidate Terminology	Body Structure	Decision
双侧甲状腺切除术	甲状腺	甲状腺病损切除术	甲状腺	retain
		双侧肾上腺切除术	肾上腺	discard
胃穿孔修补术	胃	胃修补术	胃	retain
		肠穿孔修补术	肠	discard

The DSI avoids some interference items introduced by removing supplementary information during the construction of CSTS, but it is not change the size of CSTS. The DBS discards candidate terminologies that are completely unrelated to the original terminology by comparing body structures, effectively alleviating the imbalance of positive and negative sample ratios without affecting CSTSA. Take the initial CSTS constructed by Top-30 as an example, the CSTS scale can be reduced from Top-30 to AVG-22 (namely the average quantity of the candidate terminologies contained in each CSTS is 22), which significantly reduces the proportion in the number of positive and negative samples, and reduces the computational cost of the model by 26.7%.

The BERT classification model with focal loss function. Although the DBS can effectively alleviate the problem of the imbalance of positive and negative sample ratios. But the imbalance problem still exists. The same as the Top-30 in the above paper as an example, the ratio of positive and negative samples of 1:21 after denoising is still too large. Therefore, we further use the focal loss function [16] to alleviate this problem.

The loss function used by the original BERT classification model is the cross entropy loss function, as shown in formula 2. The focal loss function is improved on the basis of the cross entropy loss, as shown in formula 3. Compared with the cross entropy loss function, the focal loss function adds two parameters α and γ. The γ is used to adjust

chinaXiv:202010.00060v2

the contribution of difficult samples to the loss function, and α to control the weight of positive and negative samples in the loss function. In the ABTSBM, we think the importance of positive and negative samples is the same, but we need to control the impact of difficult samples on the loss function, so we take the values as follows: α = 0.5, γ = 2.

$$L = -ylogy' - (1-y)log(1-y') = \begin{cases} -logy', y = 1 \\ -log(1-y'), y = 0 \end{cases} \quad (2)$$

$$L_{fl} = \begin{cases} -\alpha(1-y')^{\gamma} logy', y = 1 \\ -(1-\alpha)y'^{\gamma} log(1-y'), y = 0 \end{cases} \quad (3)$$

After using focal loss function, the model's ability for candidate terminologies that are difficult to distinguish will become stronger. As shown in Table 4, in the model using cross entropy loss function, when two candidate terminologies are too similar, the probability difference of the model output may be extremely small, which can be considered that the model cannot deal with the difficult distinction situation effectively. For example, even in the case of "右胫骨骨折闭合复位髓内钉内固定", the candidate terminology with the highest probability is the correct standard terminology, in the case of "(左侧)甲状腺腺叶(单侧)切除术", the second probability is the correct standard terminology. But in terms of probability, the candidate terminologies in these two cases are very similar, with almost no difference. After using the focal loss function, the probability difference between the two terminologies is enlarged, and then the model truly has the ability to deal with this difficult distinction situation.

Table 4. Prediction probability comparison between cross entropy loss model and focal loss model.

Original Terminology	Candidate Terminology	Probability of Positive Sample	
		Cross Entropy Model	Focal Loss Model
右胫骨骨折闭合复位髓内钉内固定	胫骨骨折闭合复位内固定术	0.9989103	0.9201928
	胫骨骨折切开复位内固定术	0.99729234	0.14210172
(左侧)甲状腺腺叶(单侧)切除术	单侧甲状腺切除伴他叶部分切除术	0.9998492	0.05657335
	单侧甲状腺叶切除术	0.9978193	0.93587375

4 Experiment and analysis

4.1 Experiments data

The experimental data comes from the ICD-9 terminology standardization academic competition organized by CHIP2019.

The CHIP2019 academic competition provides 9866 ICD-9 standard terminologies and 5492 terminology pairs (each terminology pair composes of the original terminology and its ICD-9 standard terminology), of which 3642 are taken as training dataset

chinaXiv:202010.00060v2

and the remaining are taken as test dataset. In the BTSBM and ABTSBM, if the correct standard terminology is not contained in corresponding CSTSs that are taken as training dataset, we will manually add the correct standard terminology to CSTSs.

4.2 Evaluation metrics

We use the precision defined by CHIP2019 as the final evaluation metrics, as shown in formula 4. In this task, because of the task characteristics, the final evaluation metric only consider precision. However, CSTSA is indicative for the construction of CSTS. So in the CSTS construction process, CSTSA should still be considered, the formula as shown in formula 5.

$$precision = \frac{The\ number\ of\ correct\ pairs\ of\ original\ and\ standard\ terminology}{Total\ number\ of\ original\ terminologies} \quad (4)$$

$$CTSTA = \frac{Total\ number\ of\ CSTS\ containing\ the\ correct\ standard\ terminology}{Total\ number\ of\ original\ terminologies} \quad (5)$$

4.3 Experimental results and analysis

CSTSA. We compare the CSTSA of the BTSBM after two denoising methods with the original BTSBM, as shown in Table 5. The Top-N column represents CSTS scale. The BTSBM column represents the CSTSA of the BTSBM. The BTSBM-DSI column represents CSTSA of the BTSBM after the DSI. The AVG-N (for BTSBM-DBS) column represents the average CSTS scale of the Top-N by the BTSBM after the DBS and the content in brackets is its original scale. The BTSBM-DBS column represents the CSTSA of the BTSBM-DBS. Especially, because the ABTSBM also uses DBS, the CSTSA of ABTSBM is the same as BTSBM-DBS.

Table 5. The CSTSA of BTSBM-DSI and BTSBM-DBS with BTSBM.

Top-N	BTSBM	BTSBM-DSI	AVG-N (for BTSBM-DBS)	BTSBM-DBS
Top-15	84.8	83.9	AVG-12 (Top-15)	84.8
Top-20	86.1	86.2	AVG-15 (Top-20)	86.1
Top-30	89.5	89.1	AVG-22 (Top-30)	89.5
Top-40	90.4	90.4	AVG-30 (Top-40)	90.4
Top-50	91.4	91.3	AVG-35 (Top-50)	91.3

It can be seen from Table 5 that BTSBM-DSI hardly affects CSTSA when the N value is large. And BTSBM-DBS reduced the original datasets of Top-15, Top-20, and Top-30 to average sizes AVG-12, AVG-15, and AVG-22, which significantly reduced the proportion of positive and negative samples while maintaining the original CSTSA, and also hardly affects CSTSA.

chinaXiv:202010.00060v2

Experimental results of BTSBM. There is the precision of BTSBM, which are shown in Table 6. The dataset constructed for the BERT classification model through N-gram algorithm, includes the training dataset with Top-N_1 scale CSTS and the test dataset with Top-N_2 scale CSTS:

Table 6. The precision of BTSBM.

BTSBM		
Training Dataset	Test Dataset	
	Top-20	Top-30
Top-15	79.2	81.3
Top-20	79.9	82.6
Top-30	80.0	82.9

Experimental results of ABTSBM. We compare the precision of BTSBM after DSI (namely BTSBM-DSI) and BTSBM after DBS (namely BTSBM-DBS) with original BTSBM, as shown in Table 7. And we also compare the precision of BTSBM with **F**ocal **L**oss function (namely BTSBM-FL) with original BTSBM, as shown in Table 8. Finally, we compare the precision of ABTSBM with BTSBM, as shown in Table 9.

Table 7. The precision of BTSBM-DSI and BTSBM-DBS compared with BTSBM.

Training Dataset	BTSBM		BTSBM-DSI		BTSBM-DBS		
	Test Dataset		Test Dataset		Training Dataset After DBS	Test Dataset	
	Top-20	Top-30	Top-20	Top-30		AVG-15 (Top-20)	AVG22 (Top-30)
Top-15	79.2	81.3	78.5	79.3	AVG-12(Top-15)	79.8	81.5
Top-20	79.9	82.6	78.8	80.3	AVG-15(Top-20)	80.4	82.8
Top-30	80.0	82.9	78.6	80.6	AVG-22(Top-30)	80.4	83.1

From the experimental comparison results in Table 7, it can be seen that the BTSBM-DSI does not achieve a better precision than BTSBM, though DSI removes the supplementary information to avoid candidate terminologies in CSTS that are highly similar to the original terminologies but not irrelevant. With the training dataset size set in this paper, the BTSBM can already learn features related to supplementary information to determine whether it is an irrelevant interference item. However, the DBS can reduce the training dataset with scale Top-30 to dataset with scale AVG-22. The dataset size is 26.7% of the BTSBM, which reduces the proportion of positive and negative samples. Moreover, DBS does not cause CSTSA loss. As shown in Table 7, when the training dataset and the test dataset are both Top-30, the precision of BTSBM-DBS is still improved by 0.2% compared with BTSBM, which effectively illustrates the effectiveness of the DBS method.

chinaXiv:202010.00060v2

Table 8. The precision of the BTSBM-FL compared with BTSBM.

Training Dataset	BTSBM		BTSBM-FL	
	Test Dataset		Test Dataset	
	Top-20	Top-30	Top-20	Top-30
Top-15	79.2	81.3	80.1	82.4
Top-20	79.9	82.6	80.5	83.2
Top-30	80.0	82.9	80.5	83.5

From the experimental comparison results in Table 8, it can be seen that through the BTSBM-FL, for exactly the same dataset without denoising, when the training dataset and the test dataset are both Top-30, the precision is improved by 0.6%, which fully illustrates the imbalance in the ratio of positive and negative samples does have an impact on model training, and the focal loss function can better learn and predict the unbalanced data in this task.

Table 9. The precision of the ABTSBM compared with BTSBM.

Training Dataset	BTSBM		ABTSBM		
	Test Dataset		Training Dataset After DBS	Test Dataset	
	Top-20	Top-30		AVG-15 (Top-20)	AVG-22 (Top-30)
Top-10	79.0	80.5	AVG-12(Top-15)	79.4	81.7
Top-15	79.2	81.3	AVG-15(Top-20)	80.7	83.2
Top-20	79.9	82.6	AVG-22(Top-30)	80.5	**83.5**
Top-30	80.0	82.9	-	-	-

Through the experimental comparison results in Table 9 and the horizontal comparison data, it can be seen that when the ratio of positive and negative samples between the ABTSBM and the BTSBM are almost the same, the precision of the ABTSBM is higher than the BTSBM. When the training dataset and test dataset of the ABTSBM and the BTSBM are AVG-22 (Top-30) after DBS and Top-20 without DBS respectively, the ABTSBM achieves a 0.9% higher precision than the BTSBM. Comparing the data diagonally, it can be seen that the ABTSBM not only reduces the dataset size of the original Top-30 to the dataset size of AVG-22, reducing the scale by 26.7%, which significantly reduces the computational cost and reduces proportion of positive and negative samples, but also achieved 0.6% precision higher than the BTSBM on the Top-30 test dataset.

5 Conclusion

In summary, we first propose the BTSBM that combines the BERT and text similarity. Then we subsequently propose an optimized terminology standardization method: ABTSBM, which 1) uses a large-scale initial CSTS to maintain a high CSTSA to ensure a high system performance ceiling, 2) uses the DBS to reduce the size of the CSTS

chinaXiv:202010.00060v2

without affecting CSTSA, which not only reduces the computational cost, but also reduces the imbalance of the positive and negative sample ratio of the dataset, 3) uses the BERT classification model with focal loss function to improve the model's ability to train unbalanced data by the focal loss function. Through the ABTSBM, the precision is up to 83.5%, which is 0.6% higher than BTSBM, while reducing the calculation cost by 26.7%.

References

1. Liu, T.: Development and application of clinical ICD-10 entry system. People's Military Surgeon 59(1), 96-97(2016).
2. Cheng, C., Huang, H., Ou, D.: Design and Application of Automatic Coding for Disease Diagnosis. Chinese Medical Record 19(9), 96-97(2018).
3. Nitesh, P., Manasi, G., Rajesh, W.: A Review on Text Similarity Technique used in IR and its Application. International Journal of Computer Applications, 120(2015).
4. Wang, C., Yang, Y., et al.: A Review of Text Similarity Approaches. Information Science 37(03), 158-168(2019).
5. Erjing, C., Enbo, J.: A Survey of Research on Text Similarity Calculation Methods. Data Analysis and Knowledge Discovery 1(6), 1-11(2017).
6. Brown, P. F., Pietra, V. J. D., Souza, P. V. D., et al.: Class-based n-gram models of natural language. Computational Lingus 18(4), 467-479(1992).
7. Irving, R., Fraser, C.: Two algorithms for the longest common subsequence of three (or more) strings[C]// DBLP, 1992.
8. Navarro, G.: A Guided Tour Approximate String Matching. ACM Computing Surveys (CSUR), 2001.
9. Yu, T., Xu, P., et al.: Text Similarity Method Based on the Improved Jaccard Coefficient. Computer Systems and Applications 26(12), 137-142(2017).
10. Sidorov, G., Helena, G., Markov, I., et al.: Computing Text Similarity using Tree Edit Distance[C]// Fuzzy Information Processing Society. IEEE, 2015.
11. Kenter, T., Rijke, M. D.: Short Text Similarity with Word Embeddings[C]// Acm International on Conference on Information & Knowledge Management. ACM, 2015.
12. Mikolov, T.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, 3111-3119(2013).
13. Pennington, J., Socher, R., Manning, C. D.: GloVe: Global Vectors for Word Representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014: 1532-1543.
14. Devlin, J., Chang, M. W., Lee, K., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
15. Chalapathy, R., Borzeshi, E. Z., Piccardi, M.: Bidirectional LSTM-CRF for Clinical Concept Extraction[J]. (2016).
16. Lin, T. Y., Goyal, P., Girshick, R., et al.: Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis & Machine Intelligence PP(99), 2999-3007(2017).

chinaXiv:202010.00060v2