Multi-Strategy Clinical Terminology Standardization

Authors: Lin Nankai, Lin Xiaodian, Wu Kaiying, Chen Feng, Jiang Shengyi

Date: 2023-07-11T00:00:00+00:00

Abstract

Clinical terminology standardization holds significant research importance for addressing the issue of non-standard clinical terminology in electronic medical records. Currently, the mainstream solution employs a "recall-rank" strategy. This paper proposes a multi-strategy clinical terminology standardization method based on the dataset provided in Task 3 of the China Health Information Processing Conference (CHIP2021) evaluation. In the recall stage, it employs exact matching strategy, standard term recommendation for similar original terms, and similarity calculation based on TF-IDF and improved Jaccard coefficient to recall candidate standard term sets. Simultaneously, this paper constructs a BERT-based standard term quantity prediction model, effectively improving the model's prediction performance and generalization capability through adversarial training, Focal Loss, and label smoothing strategies. In the ranking stage, it utilizes a BERT entailment score ranking model based on adversarial training and diagnostic information fusion to rank the candidate term sets, and then generates the final predicted standard terms according to the output of the quantity prediction model. In the final evaluation, the proposed method achieved an accuracy of 0.6356, ranking second among the participating teams.

Full Text

Clinical Term Normalization Based on Multiple Strategies

Nankai Lin[1], Xiaotian Lin[2], Kaiying Wu[3], Feng Chen[2], Shengyi Jiang[1],[4] [1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, 510000, China
[2]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510000, China
[3]School of Mathematics and Statistics, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510000, China

[4]Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510000, China

Abstract

Clinical term normalization is crucial for addressing the inconsistency of clinical terminology in electronic medical records. The current mainstream approach employs a "recall-and-rank" strategy. Based on the dataset provided in Evaluation Task 3 of the China Conference on Health Information Processing (CHIP 2021), this paper proposes a multi-strategy clinical term normalization method. In the recall phase, we employ exact matching, standard term recommendation from similar original terms, and similarity computation based on TF-IDF and an improved Jaccard coefficient to retrieve candidate standard term sets. Additionally, we construct a BERT-based standard term quantity prediction model, utilizing adversarial training, Focal Loss, and label smoothing to effectively enhance model performance and generalization. In the ranking phase, we use a BERT-based entailment scoring model that incorporates adversarial training and diagnostic information fusion to rank candidate terms, generating final predictions based on the output from the quantity prediction model. Our method achieved an accuracy of 0.6356 in the final evaluation, ranking second among all participating teams.

Keywords: Clinical Term Normalization; Multi-Strategy Fusion; Similarity Computation

1 Introduction

Clinical terminology standards are concept-oriented, integrated systems designed for computer applications, with disease diagnosis at their core. They cover body structures, etiology, pathology, clinical manifestations, diagnostic techniques, operative procedures, medical equipment, nursing care, social contexts, and physical factors. Intelligent diagnosis, medical imaging recognition, and other applications face challenges from non-standard clinical terminology, insufficient medical knowledge, and lexical, syntactic, semantic, and pragmatic uncertainties. A unified, standardized terminology system is essential for achieving standardized and normalized underlying data across systems, enabling precise expression of medical concepts, coding, extraction, and analysis of medical data, and supporting consistent indexing, storage, retrieval, and cross-system integration. This facilitates semantic interoperability of healthcare data and plays a vital role in medical artificial intelligence.

Semantic relationships, also known as semantic structures, abstract and generalize relationships between conceptual meanings of terms. To achieve intelligent autonomy in medical AI, systems must be capable of associating and processing semantic relationships in data. Chinese clinical terminology standards utilize similarity, suspiciousness, and deep learning algorithms to process natural language, deeply mine potential semantic relationships, and establish clinically

meaningful associations between diseases/diagnoses and anatomical sites, clinical manifestations, observations, examinations, treatments, pathology, chemicals, drugs, morphology, and other medical elements. This provides a fact-based computational context for medical AI mechanisms and supports solving practical challenges in medical AI applications. Standardized clinical terminology can eliminate uncertainty in clinical concepts, support precise recording and analysis of medical data, enable sharing and utilization of medical data across different systems, and promote deep integration between AI and healthcare.

In 2013, ShARe/CLEF eHealth [**?**] first released English clinical term normalization data. Subsequently, SemEval-2014 Task 7 [**?**] and SemEval-2015 Task 14 [**?**] released English clinical term normalization evaluation tasks. The Fifth China Conference on Health Information Processing (CHIP 2019) [**?**] introduced a Chinese clinical term normalization task and dataset, advancing research in Chinese natural language processing. However, Chinese clinical term normalization remains in its early stages with limited research, primarily due to varied expression habits among healthcare professionals. In practice, the same diagnosis, procedure, medication, test, or symptom can have hundreds or thousands of different expressions. The clinical term normalization task in medical information processing aims to find corresponding standard expressions for these varied clinical descriptions.

Early approaches to clinical term normalization relied on rule-based methods. Ghiasvand et al. [**?**] used edit distance features to generate candidate sets, learning 554 edit distance patterns from training data and achieving the best performance on SemEval-2014 Task 7. Kang et al. [**?**] proposed five rules to improve disease term normalization performance.

Currently, most clinical term normalization tasks adopt a "recall-and-rank" strategy. Leaman et al. [**?**] first proposed a pairwise learning-to-rank technique that uses a vector space model to compute textual similarity between non-standard and standard medical entities. Luo et al. [**?**] proposed a multi-task framework that can normalize disease and procedure entities simultaneously, where shared structures enable the model to leverage medical correlations between diseases and procedures to better perform disambiguation. Ji et al. [**?**] implemented entity normalization through fine-tuned pre-trained BERT models.

For Chinese clinical term normalization, Chong et al. [**?**] built a clinical term normalization system based on textual entailment, consisting of data preprocessing, BERT entailment scoring, BERT quantity prediction, and logistic regression-based reranking modules, achieving 94.825% performance and ranking first in CHIP 2019 Evaluation Task 1. Chen et al. [**?**] designed two retrieval methods to obtain candidate standard terms by searching "code-standard term" pairs and "annotation history," then reranked candidates based on textual entailment, achieving 89.1% with a single model and 92.8% with an ensemble model on the test set. Sun et al. [**?**] proposed a BERT-based clinical term normalization method that uses Jaccard similarity to select candidate terms from the stan-

dard terminology set and matches original terms with candidates using BERT, achieving 90.04% accuracy. Similar to Sun et al., Yang et al. [**?**] combined textual similarity ranking with BERT model matching, achieving 88.51% accuracy on CHIP 2019 Evaluation Task 1.

Beyond the "recall-and-rank" strategy, some researchers have explored generative approaches. Yan [**?**] analogized clinical term normalization to translation tasks, introducing deep generative models to generate core semantics of description texts and obtain standard term candidate sets, then reranking candidates using BERT-based semantic similarity algorithms.

2 Multi-Strategy Clinical Term Normalization Method

As shown in Figure 1 [Figure 1: see original paper], we propose a multi-strategy clinical term normalization method comprising three modules: candidate standard term recall, standard term quantity prediction, and candidate standard term ranking. In the recall phase, we employ exact matching, standard term recommendation from similar original terms, and similarity computation based on TF-IDF and an improved Jaccard coefficient to retrieve candidate standard term sets. Simultaneously, we construct a BERT-based standard term quantity prediction model, utilizing adversarial training, Focal Loss, and label smoothing to effectively improve prediction performance and generalization. In the ranking phase, we use a BERT-based entailment scoring model incorporating adversarial training and diagnostic information fusion to rank candidate terms, generating final predictions based on the quantity prediction model's output.

2.1 Candidate Standard Term Recall

In the recall phase, we employ three strategies: exact matching, standard term recommendation from similar original terms, and similarity computation based on TF-IDF and an improved Jaccard coefficient. Let the ICD-10 standard term list be $S = \{S_1, S_2, \ldots, S_n\}$, the training set samples be $X = \{X_1, X_2, \ldots, X_n\}$, the original term of sample X_i ($X_i \in X$) be O_i, and the candidate standard term list be H_i.

First, we apply exact matching to filter candidate standard terms. For each sample X_i, we iterate through the standard term list S and check whether each S_j ($S_j \in S$) appears as a complete substring in O_i. If so, we add it to the candidate standard term list H_i.

In the similar original term recommendation module, we train a TF-IDF model M_O on all original terms in X. Using this model, we compute cosine similarity between each sample's original term O_i and all other original terms in X. For each sample X_i, we select the top k most similar original terms and add their corresponding standard terms to H_i. In our experiments, k is set to 40.

For the TF-IDF and improved Jaccard coefficient-based similarity computation module, we train a TF-IDF model M_s on the ICD-10 standard term list S. This

model calculates cosine similarity Sim_T between sample X_i's original term O_i and each standard term S_j. Additionally, we compute character-level Jaccard similarity between O_i and S_j. The original Jaccard coefficient is:

$$Jaccard(O_i, S_j) = \frac{|O_i \cap S_j|}{|O_i \cup S_j|}$$

Since original terms may contain multiple standard terms and thus carry substantially more information than individual standard terms, the Jaccard coefficient tends to be low. Therefore, when computing the Jaccard coefficient, we only consider the number of characters appearing in the standard term in the denominator:

$$Jaccard(O_i, S_j) = \frac{|O_i \cap S_j|}{|S_j|}$$

We denote the similarity computed by this improved Jaccard coefficient as Sim_j. The fused similarity between original term O_i and standard term S_j is:

$$S = Sim_T + Sim_j$$

We compute fused similarity between O_i and each standard term in S, then select the top r standard terms by similarity and add them to H_i. In our experiments, r is set to 100.

Candidate standard terms recalled through these three steps may contain duplicates. After deduplication, we obtain the final candidate standard term list.

2.2 Standard Term Quantity Prediction

We construct a standard term quantity prediction model based on BERT (Figure 2 [Figure 2: see original paper]), incorporating various strategies to enhance prediction performance and generalization. During evaluation, we ensemble three strategy variants: BERT with adversarial training, BERT with adversarial training and Focal Loss, and BERT with adversarial training and label smoothing. The model's labels are divided into three categories: containing one standard term, containing two standard terms, and containing more than two standard terms.

2.2.1 Adversarial Training We employ adversarial training to increase model diversity and generalization, using the Fast Gradient Method (FGM) to inject noise into the model's embedding layer. The perturbation is defined as:

$$r_{adv} = \alpha \cdot \frac{g}{\|g\|_2}$$

where g is the original model gradient and α is the weight matrix for sentence vocabulary, set to 1 in our experiments. The perturbed gradient is:

$$g' = r_{adv} + g$$

During adversarial training, we perform backpropagation and parameter updates using the perturbed gradient, then remove the noise from the embedding layer to restore the original gradient for the next iteration.

2.2.2 Focal Loss Due to class imbalance where negative samples outnumber positive samples, classification results may be biased. In addition to cross-entropy loss, we adopt Focal Loss to mitigate this issue by reducing the weight of numerous easy negative samples during training. The Focal Loss formula is:

$$L = \sum (1 - p_i)^{\gamma} \log(p_i)$$

where the weight coefficient γ is a hyperparameter. Lin et al. [**?**] verified that the optimal value for γ is 2, which we adopt in our experiments.

2.2.3 Label Smoothing For cross-entropy loss, the model fits one-hot encoded labels during training, which can lead to overfitting on true labels and compromise generalization ability. We employ label smoothing to alleviate this overfitting. Let y be a one-hot encoded label; the smoothed label becomes:

$$y_i' = (1 - \varepsilon) \cdot y_i + \frac{\varepsilon}{K}$$

The smoothed loss is:

$$L = \sum y_i' \cdot \log(p_i)$$

where ε is the smoothing factor and K is the number of classes. For standard term quantity prediction, $K = 3$ and ε is set to 0.05.

2.3 Candidate Standard Term Ranking

We construct a textual entailment model based on BERT (Figure 3 [Figure 3: see original paper]) to compute entailment scores between a given diagnostic original term and each standard term. We concatenate the original term with a candidate standard term and feed them into BERT for 0-1 classification, where 0 indicates the candidate is not a correct standard term and 1 indicates it is. We use the probability of classifying as 1 as the entailment score.

We ensemble four strategy variants: BERT, BERT with adversarial training, BERT fused with surgical original term data, and BERT with both adversarial training and fused surgical original term data.

For data construction, we treat surgical original term data as positive samples only. The evaluation task provides 2,500 original term samples, which become 2,605 after processing. Due to the small sample size, we augment this data fourfold, generating 10,420 surgical original term training samples. For diagnostic original term data construction, we use the similar original term recommendation module from Section 2.1 to extract standard terms from the five most similar original terms and the ten most similar standard terms from TF-IDF model M_s. Terms not in the answer set serve as negative samples, while all standard terms in answers serve as positive samples. This strategy yields 12,984 positive samples and 89,688 negative samples. We augment positive samples sevenfold, resulting in 90,888 positive samples.

2.4 Multi-Stage Result Fusion

After obtaining recalled candidate standard terms (Section 2.1), we use the entailment model (Section 2.3) to rank them. Simultaneously, the standard term quantity prediction model (Section 2.2) predicts the number of standard terms for each test sample. If the model identifies a sample as containing one or two standard terms, we recommend the top one or two highest-scoring terms, respectively. If the predicted quantity exceeds two, we select terms with prediction probability greater than 0.5. If fewer than three terms meet this threshold, we select the top three scoring terms.

We first preprocess standard terms in the training data by removing those marked as "O" and eliminating symbols like "、". After merging all standard terms from the training set with those from the ICD-10 Beijing Clinical Version v601 and deduplicating, we obtain a new standard term list containing 37,869 terms. The distribution of preprocessed standard term quantities is shown in Figure 4 [Figure 4: see original paper], categorized as containing one, two, or more than two standard terms. We also analyzed length characteristics of diagnostic original terms, with results presented in Table 1 .

4 Experimental Results and Analysis

4.1 Experimental Setup

We conducted experiments on an RTX TITAN GPU using PyTorch 1.7.0 and Transformers 4.4.0. For both the standard term quantity prediction and candidate ranking modules, we performed five-fold cross-validation. In the candidate ranking module, surgical original term data was used for training each fold. During prediction, each model's output is the average probability from the five cross-validation models, with different model results fused by averaging their probabilities. In the recall phase, we evaluate using strict accuracy, considering

a sample correct only if all standard terms in the answer are fully recalled. Classification models in the quantity prediction and ranking modules are evaluated using accuracy. We use MedBERT-wwm[2] as the pre-trained base model for both modules, with fine-tuning parameters shown in Table 2 .

4.2 Candidate Term Recall Results

We first investigated the effectiveness of fusing TF-IDF with the improved Jaccard coefficient, testing several combinations: TF-IDF only, Jaccard only, improved Jaccard only, TF-IDF + Jaccard, and TF-IDF + improved Jaccard. Results in Table 3 show that improved Jaccard alone performs slightly worse than original Jaccard (accuracy drops 0.79%). However, when fused, TF-IDF with improved Jaccard achieves the best performance, improving over TF-IDF alone and improved Jaccard alone by 0.46% and 2.0%, respectively, demonstrating that fusion effectively enhances recall performance.

We further conducted ablation studies to validate the effectiveness of the three recall methods (Table 4). Removing exact matching decreases accuracy by 0.77%, removing similar term recommendation causes a significant 19.51% drop, and removing similarity computation results in a 21.51% decrease, confirming the contribution of each component.

4.3 Standard Term Quantity Prediction Results

We examined improvements from different strategies (Table 5). Adversarial training provides effective enhancement, increasing accuracy by 0.63%. Additionally, loss function modifications (Focal Loss and label smoothing) deliver stable performance improvements.

4.4 Candidate Standard Term Ranking Results

Results for our four textual entailment models are shown in Table 6 . The BERT model with both adversarial training and fused surgical original term data performs best, reaching 94.26% accuracy. Fusing surgical data alone causes a slight performance decrease, but we still ensemble this model's results during final evaluation to improve generalization.

4.5 Error Analysis

We analyzed errors in the candidate term recall module (Table 7). Three examples illustrate cases where full recall fails: (1) when similarity between standard and original terms is too low, and (2) when original terms correspond to too many standard terms, preventing complete recall.

In the ranking module, the model often misidentifies hypernyms of correct standard terms as the standard terms themselves. For example, for the original term left sole laceration with infection with standard term foot soft tissue infection, the model incorrectly identifies the hypernym infection as being en-

tailed. Similarly, for the original term double uterus with double cervix and double vagina with standard term double uterus with double cervix and double vagina, the model incorrectly identifies the partial term double uterus as being entailed. This hypernym misjudgment significantly impacts entailment model performance.

The quantity prediction module struggles to accurately determine the number of standard terms from original text alone, particularly in two extreme cases: long original texts corresponding to only one or two standard terms, and short original texts containing multiple standard terms (Table 8).

4.6 Final Test Set Results

In the final evaluation, we ensembled three strategy variants for quantity prediction and four variants for candidate ranking. Our method achieved an accuracy of 0.6356, ranking second among all participating teams.

5 Conclusion

For CHIP 2021 Evaluation Task 3, we propose a multi-strategy clinical term normalization method designed to enhance model generalization at each stage and improve overall performance. In the recall phase, we employ exact matching, similar original term recommendation, and similarity computation based on TF-IDF and improved Jaccard coefficient. We construct a BERT-based standard term quantity prediction model enhanced with adversarial training, Focal Loss, and label smoothing. In the ranking phase, we use a BERT entailment scoring model incorporating adversarial training and diagnostic information fusion. Our method achieved 0.6356 accuracy in the final test set, ranking second.

Our approach has limitations. We only utilized BERT as the pre-trained model and used surgical original term data solely for fine-tuning in the ranking module without fully leveraging its information. Future work will explore different pre-trained models and investigate better ways to fuse and extract information from surgical original term data to assist the task.

References

[1] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013[C]//International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, 2013: 212-231.

[2] Sameer Pradhan, Noemie Elhadad, Wendy Chapman, et al. Semeval-2014 task 7: Analysis of clinical text[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: 54-62.

[3] Noemie Elhadad, Sameer Pradhan, Sharon Gorman, et al. SemEval-2015 task 14: Analysis of clinical text[C]//proceedings of the 9th International Workshop

on Semantic Evaluation (SemEval 2015), 2015: 303-310.

[4] Huang Yuanhang, Jiao Xiaokang, Tang Buzhou, Chen Qingcai, Yan Jun. Overview of CHIP2019 Evaluation Task 1: Clinical Term Normalization Task[J]. Journal of Chinese Information Processing, 2021, 35(03): 94-99.

[5] Ghiasvand O, Kate R J. R.: UWM: Disorder mention extraction from clinical text using CRFs and normalization using learned edit distance patterns[C]//Proceedings of the 8th International Workshop on Semantic Evaluation, 2014: 828-832.

[6] Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text[J]. Journal of the American Medical Informatics Association, 2013, 20(5): 876-881.

[7] Leaman R, Islamaj Doğan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.

[8] Luo Y, Song G, Li P, et al. Multi-task medical concept normalization using multi-view convolutional neural network[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[9] Zongcheng Ji, Qiang Wei, Hua Xu. BERT-based ranking for biomedical entity normalization[J]. AMIA Summits on Translational Science Proceedings, 2020.

[10] Chong Weifeng, Li Hui, Li Xue, Ren He, Yu Dong, Wang Yehan. A Term Normalization System Based on BERT Entailment Reasoning[J]. Journal of Chinese Information Processing, 2021, 35(05): 86-90.

[11] Chen Mosha, Qiu Wei, Tan Chuanqi. A BERT-based Re-ranking Algorithm for Surgical Name Standardization[J]. Journal of Chinese Information Processing, 2021, 35(03): 88-93.

[12] Sun Yuejun, Liu Zhiqiang, Yang Zhihao, Lin Hongfei. Clinical Term Normalization Based on BERT[J]. Journal of Chinese Information Processing, 2021, 35(04): 75-82.

[13] Yang Feihong, Sun Haixia, Li Jiao. A Method for Surgical Operation Term Normalization Fusing Text Similarity and BERT Model[J]. Journal of Chinese Information Processing, 2021, 35(04): 44-50.

[14] Yan Jinghui, Xiang Lu, Zhou Yu, Sun Jian, Chen Si, Xue Chen. Application of Deep Generative Models in Clinical Term Standardization[J]. Journal of Chinese Information Processing, 2021, 35(05): 77-85.

[15] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

Note: Figure translations are in progress. See original paper for figures.