

# CODER: Knowledge infused cross-lingual medical term embedding for term normalization

Zheng Yuan, Zhengyun Zhao  
 {yuanz17,zhao-zy15}@mails.tsinghua.edu.cn  
 Tsinghua University  
 Beijing, China

Fei Wang  
 few2001@med.cornell.edu  
 Cornell University  
 New York, NY, USA

Haixia Sun, Jiao Li  
 {sun.haixia,li.jiao}@imicams.ac.cn  
 Chinese Academy of Medical Sciences / Peking Union  
 Medical College  
 Beijing, China

Sheng Yu\*  
 syu@tsinghua.edu.cn  
 Tsinghua University  
 Beijing, China

## ABSTRACT

This paper proposes CODER: **contrastive learning on knowledge graphs for cross-lingual medical term representation**. CODER is designed for **medical term normalization** by providing **close vector representations for different terms that represent the same or similar medical concepts** with cross-lingual support. We train CODER via contrastive learning on a **medical knowledge graph (KG)** named the **Unified Medical Language System [5]**, where similarities are calculated utilizing both terms and relation triplets from KG. Training with relations injects medical knowledge into embeddings and aims to provide potentially better machine learning features. We evaluate CODER in zero-shot term normalization, semantic similarity, and relation classification benchmarks, which show that CODER outperforms various state-of-the-art biomedical word embeddings, concept embeddings, and contextual embeddings. Our codes and models are available at <https://github.com/GanjinZero/CODER> standard for joint analysis.

## CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**.

## KEYWORDS

medical term normalization, cross-lingual, medical term representation, knowledge graph embedding, contrastive learning

## ACM Reference Format:

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2021. CODER: Knowledge infused cross-lingual medical term embedding for term normalization. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 11 pages.

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference'17, July 2017, Washington, DC, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

## 1 INTRODUCTION

Electronic medical records (EMRs) contain plentiful medical information that can assist various medical research, such as clinical decision support [37], **automated phenotyping** [55], and medical knowledge mining [31]. However, analyzing the free text in EMRs remains challenging because the same medical concept can appear in **various nonstandard names**, abbreviations, and misspellings, which all need to be **normalized** to their corresponding standard terms or their concept IDs of an existing terminology system, such as the Unified Medical Language System (UMLS) [5] or SNOMED Clinical Terms (SNOMED-CT) [13]. Normalization is also important for structured EMR data. Different medical institutes use different coding systems for laboratory examinations, procedures, and medications. Thus, when conducting multi-institutional studies, normalization can facilitate mapping the database fields to the same

Medical term normalization can be regarded as either **classification or ranking problem**. Neural-network based techniques encode terms into hidden representations by recurrent neural networks or pre-trained language models (PLM). A softmax layer on the hidden representations [29, 35, 39, 50] is used for classification. **Classification methods** achieve **state-of-the-art scores** in benchmark tests but are **limited to a small set of target concepts**. In practice, terminology systems in use, such as the UMLS, have millions of concepts. Normalizing medical terms to these systems remains challenging for the classification approach due to the **enormous number of classes and the inflexibility to add new classes**. **Another approach** is to **rank candidate concepts by similarity to the term** [25, 28, 49]. It differs from the previous approach in that it is **trained as a binary classifier**, where terms and their corresponding concepts' names (target terms) form positive samples, and terms and non-corresponding concepts' names form negative samples. The scalar output from the classifier is then used as a measure of similarity to rank the candidate concepts for normalization. Similar to the classification approach, these models are **trained on labeled datasets with limited target concepts**.

**To achieve** term normalization on **very large scale and flexible** concept sets, we propose CODER based on PLM with KG contrastive learning. CODER generates embeddings for terms and ranks

the candidate concepts by the cosine similarity between the embeddings of the candidate target terms and that of the input term, avoiding training a classification model. Existing medical embeddings can be classified into word embeddings [9, 30, 38], concept embeddings [3, 6, 10, 11, 52, 56], and contextual embeddings [2, 4, 17, 20, 21, 26, 42, 56]. Word and concept embeddings have good base performances on evaluating similarity, but face the out-of-vocabulary (OOV) problem and cannot handle misspellings that are commonly present in clinical text. PLM-based contextual embeddings can relieve the OOV problems by using sub-word tokenization, but they have weaker performance than word and concept embeddings on evaluating similarity if not fine-tuned [46].

Contrastive learning is a popular technique in computer vision (CV) [8, 18] as a solution to learning representations with a very large number of classes. In each training step, instead of learning to map samples to their labels, contrastive learning aims to learn sample representations to distinguish samples of the same label (positive pairs) from those of different labels (negative pairs). For term normalization, terms of the same and different concepts naturally serve as positive and negative pairs in contrastive learning, which has been explored by SapBERT [32] using the UMLS terminology. However, as a KG, the UMLS contains more knowledge than just terms, and knowledge can serve as useful information for normalization. For example, when normalizing “poisoned by eating pufferfish” we may want “food poisoning” to rank among the top because pufferfish is a kind of food. For another example, it would be more meaningful to have the embedding of “rheumatoid arthritis” closer to “osteoarthritis” than “rheumatoid pleuritis” because both “rheumatoid arthritis” and “osteoarthritis” are subtypes of arthritis, which may affect the same body parts and share treatments. The reasoning in these examples illustrates that relational knowledge is essential to achieve meaningful medical embeddings. Therefore, we propose dual contrastive learning on both terms and relation triplets of KGs. The term-relation-term similarity are measured between a term-relation  $(h, r)$  and a term  $(t)$ . Positive and negative term-relation-term pairs are determined by KGs. Online hard pairs mining is applied to find informative pairs which are often used in person re-identification (ReID) [48]. ReID and medical term normalization are similar tasks since they both have many categories and sparse samples for each category.

To support cross-lingual term normalization, we use mBERT [12] which encodes texts in different languages to one unified space. Previous methods are usually translation-based and rely on biomedical parallel corpora [1, 43, 47]. The UMLS naturally contains massive cross-lingual medical synonyms which have not been fully utilized.

The contributions of this paper include:

- We propose a KG contrastive learning model for term normalization that uses both synonyms and relations from the UMLS to direct the generation of medical term embeddings. The learning strategy can also be applied to other KGs.
- We perform evaluations on CODER against notable existing medical embeddings. CODER achieves state-of-the-art results in zero-shot term normalization, medical concept similarity measure, and concept relation classification tasks. CODER

embeddings can be used for embedding-based term normalization directly or features for machine learning, and CODER can be fine-tuned like other PLMs.

- CODER is the first cross-lingual medical term representation with support for English, Czech, French, German, Italian, Japanese, Portuguese, Russian, Spanish, Dutch, and Chinese.

## 2 RELATED WORK

In this section, we briefly introduce the medical term normalization task, existing medical embedding, and contrastive learning.

### 2.1 Medical term normalization

Classification methods generate hidden representations of terms and use softmax layer to classify terms into concepts. Terms are encoded by CNN, RNN, or PLM [29, 35, 50]. The attention mechanism is introduced to capture important words or characters for classification [39, 50]. These classification methods only consider target terms as a category and do not utilize semantic information from target terms.

Ranking methods rank the similarity between input term and candidate target terms by training them as positive and negative pairs. DNorm [25] learns to rank target terms via calculating similarities between TF-IDF vectors. Li et al. [28] encodes input and target terms by CNN and uses a pairwise approach for ranking. NSEEN [14] trains a siamese LSTM network and uses hard negative samplings for finding informative pairs. BNE [45] encodes terms, concepts, and contexts separately, and trains the model based on term-term, term-concept, and term-context similarities. Pattisapu et al. [41] calculate graph embeddings for the target terms and rank them by the cosine distance. BIOSYN [49] uses both TF-IDF and BioBERT [26] to represent terms. Synonym marginalization is used for maximizing similarities between synonyms.

Non-English medical term normalization has also attracted the interest of researchers. Niu et al. [39] constructed the Chinese Medical Concepts Normalization dataset. Translation-based methods have been used for French, Spanish, Dutch, and German [1, 43, 47].

### 2.2 Medical embeddings

Word embeddings are typically trained by the word2vec model [36] using medical records or papers [9, 30, 38].

Concept embeddings map medical concepts to embedding vectors and cannot perform term normalization intrinsically. Sequences of medical concepts identified from EMRs or biomedical papers are considered as sentences to train embeddings with word2vec [6, 10, 11, 52]. Cui2vec [3] factorizes the CUI-CUI pointwise mutual information matrix to obtain embeddings. Zhang et al. [56] proposed Conceptual-Contextual embedding trained on both biomedical corpora and UMLS relations.

Contextual embeddings can generate different embeddings according to the context. The popularity of contextual embeddings stems from large-scale PLM like BERT [12] and ELMo [44]. Several medical contextual embeddings have been trained on different medical corpora under the architecture of BERT [2, 4, 17, 20, 26, 42]. Jin et al. [21] proposed BioELMo by training on PubMed. SapBERT [32] conducts self-alignment pre-training on UMLS synonyms for

term embeddings. Contextual embeddings have great potential for various kinds of downstream NLP tasks but typically require supervised fine-tuning. We found most existing contextual embeddings perform worse on zero-shot term normalization than word embeddings since the [CLS] token of masked language models may not reflect semantic similarity directly [27].

### 2.3 Contrastive learning

Contrastive learning has been successful in learning representations from the comparison between positive and negative pairs, especially in the field of CV [24]. SimCLR [8], MoCo [18] and SwAV [7] have achieved comparable results with supervised methods in the ImageNet dataset. Data augmentation is required for self-supervised contrastive learning to generate positive pairs. Crop, Gaussian blur, and color distortion [8] can be used in the CV domain. For NLP, DeCLUTR [16] collects positive sentence pairs from the same document. Back-translation is applied in CERT [15] to create similar sentence pairs. SapBERT [32] utilizes synonyms from the UMLS as positive pairs. Our model fully utilizes the UMLS as a KG and trains on both synonyms and relation triplets. Loss functions are another essential components for contrastive learning to maximize similarities between positive pairs and minimize similarities between negative pairs. Triplet loss [19] minimizes the relative distance between one positive and one negative pair. InfoNCE [40] considers all other samples in a mini-batch as negative pairs.

## 3 METHOD

In this section, we introduce the model architecture of CODER. We first define the embedding-based term normalization task and how we embed terms. Then we introduce the KG contrastive learning for term normalization.

### 3.1 Embedding-based Term Normalization

Let  $\mathcal{D} = \{n_i\}_{i=1}^{\|\mathcal{D}\|}$  be a concept dictionary where  $n_i$  is a concept. Each concept  $n_i$  has several terms  $\{s_i^j\}_{j=1}^{c_i}$  as synonyms. We define the embedding assigned to term  $s_i^j$  as  $\mathbf{e}_i^j \in \mathbb{R}^l$ . For an input string  $s$  with embedding  $\mathbf{e}$ , the goal is to predict its concept  $n$ :

$$\hat{n} = n_{\arg \max_i (\cos(\mathbf{e}_i^j, \mathbf{e}))},$$

where  $\cos$  is the cosine similarity measure in Euclid space.

### 3.2 Term representation

CODER takes an input term  $s$  and outputs an embedding vector  $\mathbf{e}$ . The input term  $s$  is tokenized to **sub-words** [CLS],  $s_0, s_1, \dots, s_k$ , [SEP] with two special tokens. A pre-trained language model PLM encodes  $s$  to a series of hidden states  $\mathbf{h}_i$ :

$$\mathbf{h}_{[\text{CLS}]}, \mathbf{h}_0, \dots, \mathbf{h}_k, \mathbf{h}_{[\text{SEP}]} = \text{PLM}([\text{CLS}], s_0, s_1, \dots, s_k, [\text{SEP}]).$$

We initialize the PLM by PubMedBERT [17] or mBERT [12]. The embedding of  $s$  is taken from the representation of the [CLS] token:

$$\mathbf{e} = \mathbf{h}_{[\text{CLS}]}.$$

We can also embed the term by average pooling:

$$\mathbf{e}_{\text{avg}} = \text{AvgPool}(\mathbf{h}_{[\text{CLS}]}, \mathbf{h}_0, \dots, \mathbf{h}_k, \mathbf{h}_{[\text{SEP}]}).$$

### 3.3 The contrastive learning framework for KG with synonyms

CODER learns term representations by maximizing similarities between positive term-term pairs and term-relation-term pairs from a KG. The KG records relation triplets  $\{(h, r, t)\}$ , where the head concept  $h$  and the tail concept  $t$  are from concept dictionary  $\mathcal{D}$ , and  $r \in \mathcal{R}$  is the relation between them. At each training step, we sample a batch of relation triplets  $\{(h_i, r_i, t_i)\}_{i=1}^k$  from the KG. For convenience of notation, denote  $n_i = h_i, n_{k+i} = t_i, \forall 1 \leq i \leq k$ . We sample term  $s_i$  for each  $n_i$  from  $\{s_i^j\}_{j=1}^{c_i}$  in  $\mathcal{D}$ . CODER embeds terms  $\{s_i\}_{i=1}^{2k}$  into  $\{\mathbf{e}_i\}_{i=1}^{2k}$ .

We define the term-term pair label  $\tau_{ij}$  in a mini-batch based on the concepts:

$$\tau_{ij} = \begin{cases} 1, & n_i = n_j, \\ 0, & n_i \neq n_j. \end{cases}$$

The similarity between two terms in a mini-batch is defined as:

$$S_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j).$$

A series of pair-based loss functions, such as the InfoNCE [40] have been proposed for learning representations in contrastive learning. We adopt the Multi-Similarity loss (MS-loss) [53] to learn embeddings, which incorporates hard pair mining in its formulation to accelerate the convergence. Other pair-based losses can also be used as the loss function, and we leave the comparative study of loss functions to future work. The MS-loss mines hard pairs according to positive relative similarity. For an anchor  $i$ , it defines:

$$\mathcal{N}_i := \{j | \tau_{ij} = 0, S_{ij} > \min_{\tau_{ik}=1} S_{ik} - \epsilon\},$$

$$\mathcal{P}_i := \{j | \tau_{ij} = 1, S_{ij} < \max_{\tau_{ik}=0} S_{ik} + \epsilon\},$$

where  $\epsilon$  is a margin. The MS-loss of a mini-batch is:

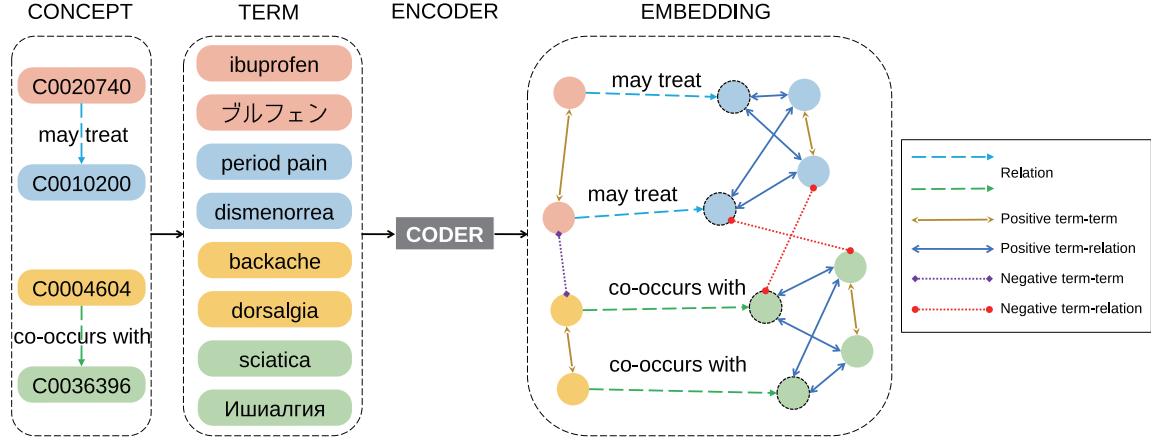
$$\mathcal{L}_{MS} = \frac{1}{2k} \sum_{i=1}^{2k} \left( \frac{\log(1 + \sum_{j \in \mathcal{P}_i} \exp(-\alpha(S_{ij} - \lambda)))}{\alpha} + \frac{\log(1 + \sum_{j \in \mathcal{N}_i} \exp(\beta(S_{ij} - \lambda)))}{\beta} \right),$$

where  $\alpha, \beta, \lambda$  are hyper-parameters.  $\mathcal{P}_i$  will vanish if  $\tau_{ij} = 0, \forall j \neq i$ , and  $\mathcal{N}_i$  will vanish if  $\tau_{ij} = 1, \forall j \neq i$ .

Toward injecting relational knowledge from KG into term representations, CODER also learns KG embedding inspired by semantic matching methods like DistMult [54] and ANALOGY [33], which aim to approximate  $M_r^T h \approx t$ . The motivation is to learn a better similarity function via relations. Consider two triplets  $(h_0, r, t_0)$  and  $(h_1, r, t_1)$  with the same relation type,  $h_0$  and  $h_1$  being semantically similar (e.g. similar diseases) may suggest that  $t_0$  and  $t_1$  are also semantically similar (e.g. similar drugs).  $M_r^T h \approx t$  will guarantee the relation-based similarity inference in the above example.  $(h_i, r_i)$  can be viewed as a relation-based data augmentation of  $t_i$ . We can define the term-relation-term similarity between relation  $(s_i, r_i), \forall 1 \leq i \leq k$  and term  $s_j, \forall k+1 \leq j \leq 2k$  in a mini-batch by:

$$S_{ij}^{\text{rel}} = \cos(\mathbf{M}_{r_i}^T \mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i^T \mathbf{M}_{r_i} \mathbf{e}_j}{\|\mathbf{M}_{r_i}^T \mathbf{e}_i\| \|\mathbf{e}_j\|}.$$

where  $\mathbf{M}_{r_i} \in \mathbb{R}^{l \times l}$  and  $\|\cdot\|$  is the L2-norm. Instead of using  $\mathbf{e}_i^T \mathbf{M}_{r_i} \mathbf{e}_j$  as the score function, CODER uses  $\cos$  to bound  $S_{ij}^{\text{rel}}$ . When  $\tau_{ij} = 1$ ,



**Figure 1: The overview of CODER.** CODER encodes terms potentially in different languages into the embedding space. Term-term similarities and term-relation-term similarities are calculated to train CODER.

$S_{ij}$  can be considered as a special case of  $S_{ij}^{rel}$  with  $\mathbf{M}_{r_i} = \mathbf{I}$ . The original DistMult samples corrupted triplets  $(h', r, t')$  as negative training sample of  $(h, r, t)$ , while CODER utilizes  $(h_i, r_i, t_{j-k})$  as negative term-relation-term pairs in a mini-batch (recall that  $k + 1 \leq j \leq 2k$  and that the index for tail entities is 1 to  $k$ ).

The term-relation-term pair label  $\tau_{ij}^{rel}$  is defined by:

$$\tau_{ij}^{rel} = \begin{cases} 1, & (h_i, r_i, t_{j-k}) \in \text{KG}, \\ 0, & (h_i, r_i, t_{j-k}) \notin \text{KG}. \end{cases}$$

It is time-consuming to check if a relation triplet belongs to the KG. In practice, we calculate:

$$\tau_{ij}^{rel} = \begin{cases} 1, & t_i = t_{j-k}, \\ 0, & t_i \neq t_{j-k}. \end{cases}$$

With  $S_{ij}^{rel}$  and  $\tau_{ij}^{rel}$  defined, the MS-loss can also be used for learning knowledge embeddings. CODER mines hard term-relation-term pairs for anchor  $(s_i, r_i)$  via:

$$\mathcal{N}_i^{rel} := \{j | \tau_{ij}^{rel} = 0, S_{ij}^{rel} > \min_{\tau_{ik}^{rel}=1} S_{ik}^{rel} - \epsilon^{rel}\},$$

$$\mathcal{P}_i^{rel} := \{j | \tau_{ij}^{rel} = 1, S_{ij}^{rel} < \max_{\tau_{ik}^{rel}=0} S_{ik}^{rel} + \epsilon^{rel}\},$$

where  $\epsilon^{rel}$  is a margin. The MS-loss of term-relation-term pairs is:

$$\mathcal{L}_{MS}^{rel} = \frac{1}{k} \sum_{i=1}^k \left( \frac{\log(1 + \sum_{j \in \mathcal{P}_i^{rel}} \exp(-\alpha^{rel}(S_{ij}^{rel} - \lambda^{rel})))}{\alpha^{rel}} + \frac{\log(1 + \sum_{j \in \mathcal{N}_i^{rel}} \exp(\beta^{rel}(S_{ij}^{rel} - \lambda^{rel})))}{\beta^{rel}} \right),$$

where  $\alpha^{rel}, \beta^{rel}, \lambda^{rel}$  are hyper-parameters. To avoid degeneration of  $\mathcal{P}_i, \mathcal{N}_i, \mathcal{P}_i^{rel}, \mathcal{N}_i^{rel}$ , we guarantee for each triplet  $(h_i, r_i, t_i)$ , there exists at least  $m$  different indexes  $\{i_o\}_{o=1}^m$  such that:

$$(h_i, r_i, t_i) = (h_{i_o}, r_{i_o}, t_{i_o}),$$

where  $2 \leq m \leq \sqrt{k}$ .

The total loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{MS} + \mu \mathcal{L}_{MS}^{rel}.$$

The parameters of CODER are updated to minimize  $\mathcal{L}$ . The overall model structure of CODER is illustrated in Figure 1.

## 4 EXPERIMENTS

In this section, we first introduce the UMLS and the training of CODER. Then we introduce the compared embeddings and the benchmarks: medical term normalization tasks, a medical similarity measure, and relation classification on concepts.

### 4.1 UMLS Metathesaurus

We leverage the UMLS to build the training dataset. The UMLS Metathesaurus contains medical concepts integrated from different lexicon resources. Each concept has a Concept Unique Identifier (CUI) with multiple synonymous names (terms) potentially in multiple languages. For example, concept **C0002502** has terms: Amilorid, Amiloride, producto con amilorida, etc. Each concept has been assigned one or occasionally multiple semantic types. For example, the semantic types of **C0002502** are Organic Chemical and Pharmacologic Substance. The UMLS also provides relation information between medical concepts in the form of triplets, such as (**C5190625**, **C0002502**, (“CHD”, “is a”)). “CHD” is the relation type and “is a” is the detailed relationship attribute in the UMLS. Some triplets only contain the relation type or the relationship attribute. We concatenate the names of the relation type and the attribute as relation labels of triplets. CODER is trained using the UMLS 2020AA release, which contains 4.27 M concepts, 15.48 M terms, and 87.89 M relations with 127 semantic types, 14 relations types, and 923 relationship attributes. Training samples are generated by relations and terms sampling according to 3.3.

## 4.2 Training settings of CODER

We train CODER<sub>ENG</sub> initialized from PubMedBERT for the English version with 100K training steps. We also train cross-lingual version CODER<sub>ALL</sub> initialized from mBERT with 1M training steps. Terms are encoded by [CLS] representation. A batch-size of  $k = 128$  relation triplets and gradient accumulation steps of 8 are used for training. In each mini-batch, we set the count of repeat triplets  $m = 8$ . The maximal sequence length is 32 since the terms are short. We use AdamW [34] as the optimizer with a linear warm-up in the first 10000 steps to a peak of  $2e-5$  learning rate that decayed to zero linearly. The parameters  $\alpha = \alpha^{rel}$ ,  $\beta = \beta^{rel}$ ,  $\lambda = \lambda^{rel}$ ,  $\epsilon = \epsilon^{rel}$  of MS-loss are set to 2, 50, 0.5, 0.1 respectively. The weight  $\mu$  is set to 1 to balance term-term loss and term-relation-term loss.

## 4.3 Compared baseline embeddings

We compare CODER with the following medical embeddings. Details of compared embeddings are listed in **Appendix A.1 Summary of compared medical embeddings**.

**Word Embedding** GoogleNews-vectors [36] is a baseline of word embedding. Wiki-pubmed-PMC [38], BioNLP-2016-win-2 [9], and BioNLP-2016-win-30 [9] are medical word embeddings with different training settings. To represent a term, we used the average of the embeddings of its words.

**Concept Embedding** Cui2vec [3], DeVine-etal-200 [52], and claims-codes-hs-300 [11] map different sets of medical concepts to vectors. Cui2vec contains medical concepts from SNOMED-CT. DeVine-etal-200 includes medical concepts that appeared in Med-Track and OHSUMED. Claims-codes-hs-300 contains ICD-9 codes, LOINC laboratory codes, and NDC medication codes.

**Contextual Embedding** BERT-base-cased and mBERT [12] are the baselines of contextual embedding. BioBERT [26] is the most well-known biomedical language model. There exist two clinical language models that have the same name ClinicalBERT [2, 20], and we distinguish them by the author's name. We compare SciBERT [4], BlueBERT [42] and PubMedBERT [17] which are biomedical specific language models. We also compare SapBERT [32], which is a BERT-based medical term representation trained using contrastive learning on the terms of the UMLS. We calculate two representations for contextual embedding  $e$  and  $e_{avg}$  for a term as we described in Section 3.2.

## 4.4 Medical term normalization tasks

We evaluate term normalization on three datasets in different languages. Term normalization tasks are evaluated in a zero-shot setting, i.e., no training datasets are provided.

**4.4.1 Cadec, PsyTar.** Cadec [22] and PsyTar [57] are two English medical term normalization datasets. Cadec contains 6754 terms and 1029 standard terms as normalization targets. PsyTar contains 6556 terms, which are mapped to 618 standard terms. We use the data splittings from [35, 50] and top- $k$  accuracy as the metric. We report the average scores on different test sets.

Results on Cadec and PsyTar are shown in Table 1. Appendix A.2 shows the results of contextual embeddings using [CLS] representation. For both datasets, CODER achieved superior performance over other baseline embeddings. CODER<sub>ENG</sub> based on KG

contrastive learning outperformed SapBERT that only uses terms by 5.72, 4.82, 1.47, 5.08 in terms of Cadec acc@1, acc@3, PsyTar acc@1, and acc@3, respectively. CODER<sub>ALL</sub> is trained by multiple languages and this may have caused CODER<sub>ALL</sub> to perform slightly worse than CODER<sub>ENG</sub>. We observe two interesting phenomena: (1) Contextual embeddings pre-trained by masked language models fail to achieve embedding-based normalization. (2) Medical word embeddings do not improve the word embedding baseline.

**Table 1: Acc@k for different embeddings in Cadec and PsyTar datasets. Contextual embeddings report results using the average pooling representation.**

Embedding	Cadec		PsyTar	
	acc@1	acc@3	acc@1	acc@3
GoogleNews-vectors	39.15	52.41	29.75	47.52
Wiki-pubmed-PMC	35.04	47.14	23.97	36.02
BioNLP-2016-win-2	35.47	48.30	25.29	37.70
BioNLP-2016-win-30	38.78	51.64	28.27	43.26
BERT-base-cased	19.64	25.65	15.16	19.84
BioBERT	3.62	17.62	7.64	10.98
ClinicalBERT(Huang)	17.04	23.25	13.97	19.40
ClinicalBERT	13.77	18.40	11.12	15.47
SciBERT	12.50	17.95	11.04	15.12
BlueBERT	14.64	20.62	13.24	18.07
PubMedBERT	11.50	15.88	9.76	13.78
SapBERT	54.05	71.42	52.45	66.04
CODER <sub>ENG</sub>	<b>59.77</b>	<b>76.24</b>	<b>53.92</b>	<b>71.12</b>
CODER <sub>ALL</sub>	56.02	72.11	43.86	60.20

**4.4.2 The MANTRA GSC.** The MANTRA GSC [23] contains French, German, Spanish, and Dutch corpus and includes medical terms mapped to a subset of the UMLS from three different corpora (MEDLINE, EMEA, and Patent). To our knowledge, only part of these non-English corpora have been evaluated on before. We compare CODER<sub>ALL</sub> with two translation-based methods: biomedical translation model (BTM) [47] and a combination of Google Translator and Bing Translator (GB) [47]. We also compare with mBERT as a multi-lingual contextual embedding baseline.

CODER<sub>ALL</sub> and mBERT search standard terms in SNOMED-CT, MeSH, MedDRA of UMLS2020AA including 801038 concepts and return one CUI per input term according to the cosine similarity. We use F1-score as the metric since one term can map to multiple CUIs in the MANTRA GSC.

Table 2 summarizes the results of MANTRA. CODER<sub>ALL</sub> benefits from the UMLS multi-language synonyms, and has made obvious improvement over mBERT. BTM is trained using additional dictionary and GB are trained on massive parallel corpus and achieve better performance than CODER<sub>ALL</sub> in the French and Dutch MEDLINE. CODER<sub>ALL</sub> achieves state-of-the-art performance in the Spanish and Germany MEDLINE, EMEA, and Patent corpora.



**Table 2: F1-score of the MANTRA GSC for translation and embedding-based methods.**

Method	SPA	FRE	DUT	GER
MEDLINE				
BTM	69.1	67.4	61.4	66.3
GB	68.7	<b>68.6</b>	<b>64.8</b>	67.9
mBERT(e)	13.6	16.7	16.3	17.3
mBERT(e <sub>avg</sub> )	20.6	19.8	17.1	21.4
CODER <sub>ALL</sub> (e)	70.1	58.6	58.6	<b>69.0</b>
CODER <sub>ALL</sub> (e <sub>avg</sub> )	<b>70.4</b>	58.9	59.5	68.4
EMEA				
mBERT(e)	17.9	25.5	16.3	14.3
mBERT(e <sub>avg</sub> )	22.2	29.0	20.1	18.4
CODER <sub>ALL</sub> (e)	<b>68.1</b>	<b>62.9</b>	61.7	65.3
CODER <sub>ALL</sub> (e <sub>avg</sub> )	<b>68.1</b>	61.1	<b>62.9</b>	<b>65.6</b>
Patent				
mBERT(e)	-	25.6	-	19.8
mBERT(e <sub>avg</sub> )	-	32.4	-	24.3
CODER <sub>ALL</sub> (e)	-	70.8	-	69.0
CODER <sub>ALL</sub> (e <sub>avg</sub> )	-	<b>71.1</b>	-	<b>70.2</b>

#### 4.5 MCSM

Choi et al. [11] introduce Medical Conceptual Similarity Measure (MCSM) for evaluating medical embeddings. MCSM is defined by:

$$\text{MCSM}(V, T, k) = \frac{1}{|V(T)|} \sum_{v \in V(T)} \sum_{i=1}^k \frac{1_T(v(i))}{\log_2(i+1)},$$

where  $V$  is a set of medical concepts,  $T$  is a semantic type,  $|V(T)|$  is the number of medical concepts with semantic type  $T$ ,  $k$  is neighborhood size,  $v(i)$  is the  $i^{\text{th}}$  closest neighbor, and  $1_T(\cdot)$  is the indicator function of semantic type  $T$ . We re-implement the evaluation code of MCSM to support word and contextual embeddings.

We use SNOMED-CT as  $V$  which contains 157726 concepts. Following the setting of [11], Pharmacologic Substance (PS), Disease or Syndrome (DS), Neoplastic Process (NP), Clinical Drug (CD), Finding (FD), and Injury or Poisoning (IP) are evaluated semantic types  $T$  and  $k$  is set to 40.

MCSM results are reported in Table 3, where terms are represented by average pooling for contextual embeddings. Appendix A.3 lists MCSM results using [CLS] representation. Concept embeddings suffer from OOV, which cannot cover all SNOMED-CT concepts and perform poorly on MCSM. **Though CODER is not designed for semantic classification or clustering, it still learns semantic information via relations from the UMLS and achieves state-of-the-art in five categories.** CODER is only bettered by word embeddings with a small margin in IP.

#### 4.6 Diseases Database relation classification

We introduce the Diseases Database relation classification (DDBRC) task to classify the relation type using the head and tail concept embeddings. 131,741 medical relation triplets are collected from

**Table 3: MCSM scores for different embeddings. Contextual embeddings report results with average pooling representation.**

Embedding	PS	DS	NP	CD	FD	IP
cui2vec	4.2	7.2	7.6	4.3	7.2	8.4
DeVine-etal-200	7.3	6.9	6.3	1.2	4.6	2.5
claims-codes-hs-300	9.4	7.7	7.5	-	3.4	8.6
GoogleNews-vectors	9.1	9.3	9.6	10.0	8.5	10.3
Wiki-pubmed-PMC	8.6	9.4	10.0	9.8	8.9	10.3
BioNLP-2016-win-2	9.1	9.4	10.0	10.2	8.9	<b>10.4</b>
BioNLP-2016-win-30	8.3	9.3	10.0	9.6	8.8	10.3
BERT-base-cased	8.6	8.4	8.7	9.7	8.4	9.3
BioBERT	8.7	8.8	9.3	9.5	7.8	9.3
ClinicalBERT(Huang)	8.3	7.5	7.8	9.7	7.5	9.3
ClinicalBERT	8.4	8.5	8.9	9.6	7.5	9.6
SciBERT	7.6	6.9	6.2	8.5	6.3	7.5
BlueBERT	8.1	7.4	8.5	9.3	7.4	9.2
PubMedBERT	9.4	9.4	9.5	10.0	8.7	10.2
SapBERT	8.2	8.3	8.5	7.9	8.6	9.1
CODER <sub>ENG</sub>	<b>10.2</b>	9.4	9.4	<b>10.7</b>	9.6	10.1
CODER <sub>ALL</sub>	<b>10.2</b>	<b>9.5</b>	<b>10.1</b>	10.6	<b>9.8</b>	10.3

the Diseases Database <sup>1</sup>. A triplet contains a head concept, a tail concept, and a relationship that can be classified into 14 categories. 18,419 of the head and tail concept pairs overlapped with the UMLS (though the concepts may have different recorded relation types). We split the data 4:1 randomly as training and test sets.

Different embeddings are evaluated with the same classification model. Denote the embedding of the head concept as  $\mathbf{e}_h$  and that of the tail concept as  $\mathbf{e}_t$ . The relation is predicted by:

$$\hat{y} = \text{softmax}(W[\mathbf{e}_h, \mathbf{e}_t] + \mathbf{b}).$$

We use cross-entropy as the loss function. Embeddings are tested as fixed (feature-based) or trainable (fine-tuned). Further training details are listed in Appendix A.4. The classification accuracy on the test set is used as the metric.

Results of DDBRC are listed in Table 4. Contextual embeddings are powerful when fine-tuned. Whether embeddings are trainable or not, CODER always achieves the highest accuracy. Under the feature-based setting, CODER<sub>ENG</sub> improves 7.1, 4.05, 4.25 over the best of concept, word, and contextual embeddings, respectively. CODER also has the minimum performance difference between feature-based and fine-tuned among all contextual embeddings, and this indicates that the medical knowledge encoded into CODER makes them informative machine learning features.

<sup>1</sup><http://www.diseasesdatabase.com/>

**Table 4: Testing accuracy of DDBRC for training 50 epochs.**

Embedding	feature	fine-tune
cui2vec	88.95	88.92
DeVine-etal-200	80.42	80.44
claims-codes-hs-300	72.02	71.99
GoogleNews-vectors	90.96	90.85
Wiki-pubmed-PMC	91.67	91.16
BioNLP-2016-win-2	91.94	92.00
BioNLP-2016-win-30	92.00	91.46
BERT-base-cased	78.22	98.74
BioBERT	83.47	98.73
ClinicalBERT(Huang)	75.82	98.77
ClinicalBERT	76.12	98.76
SciBERT	77.79	98.80
BlueBERT	79.70	98.78
PubMedBERT	75.93	98.81
SapBERT	91.80	98.79
CODER <sub>ENG</sub>	<b>96.05</b>	<b>98.85</b>
CODER <sub>ALL</sub>	95.82	98.81

## 5 DISCUSSIONS

### 5.1 Ablation study

To explore how term-relation-term similarity impacts the normalization performance, we first train CODER without term-relation-term similarity. We further test another TransE-like term-relation-term similarity function  $S_{ij}^{rel,T} = -\|\mathbf{e}_i + \mathbf{r}_i - \mathbf{e}_j\|$ . Details of pre-training parameters are listed in Appendix A.5.

**Table 5: Ablation results for different training settings of CODER.**

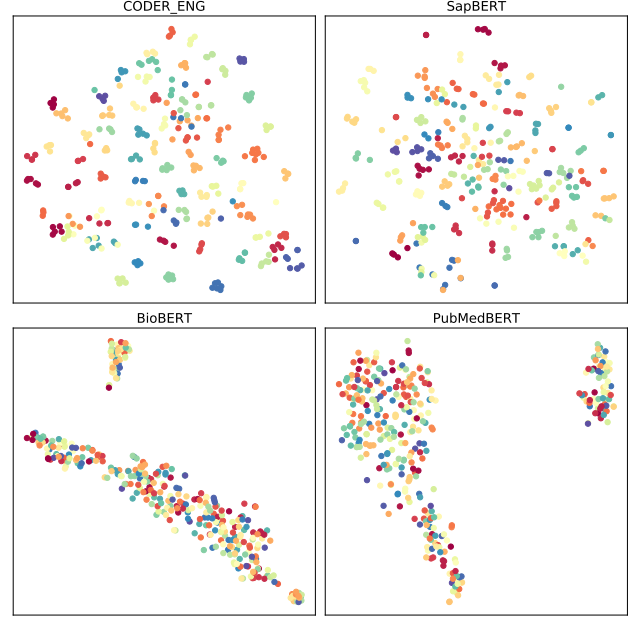
Setting	Cadec		PsyTar	
	acc@1	acc@3	acc@1	acc@3
CODER <sub>ENG</sub>	<b>59.77</b>	<b>76.24</b>	<b>53.92</b>	<b>71.12</b>
$-S_{ij}^{rel}$	54.18	71.40	50.58	67.22
$-S_{ij}^{rel} + S_{ij}^{rel,T}$	57.35	73.61	51.35	69.23

Table 5 shows ablation results on dataset Cadec and PsyTar. Term normalization performance gains from relation information of the UMLS.  $S_{ij}^{rel,T}$  utilizes relation information of the UMLS, but it has an unbounded range and is inconsistent with the form of  $S_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$ . Therefore, using  $S_{ij}^{rel}$  is a better choice to encode relational knowledge with better normalization performance.

### 5.2 Embedding visualization analysis

We use t-SNE [51] to visualize different medical embeddings. Figure 2 and Figure 3 plot medical terms from the International Classification of Diseases ICD-10-CM. Terms attached to the same concept are in the same color for Figure 2 and concepts are color-coded by semantic types in Figure 3. They show that CODER<sub>ENG</sub> and SapBERT cluster terms automatically and group concepts according to

semantic types, while BioBERT and PubMedBERT are incapable of clustering. Notice that CODER<sub>ENG</sub> can distinguish concepts more clearly. SapBERT mixes up Sign or Symptom, Disease or Syndrome, and Injury or Poisoning a little, while CODER<sub>ENG</sub> achieves better category clustering.

**Figure 2: The t-SNE Visualization of different contextual embeddings for term normalization.**

### 5.3 Case study of CODER

**5.3.1 Cross-lingual.** To illustrate the ability of CODER<sub>ALL</sub> on cross-lingual term normalization, we search the closest English term to the terms in different languages for concept C0004604 “*dorsalgia*” as a case study, as there is no medical normalization dataset for multiple languages. We search for the most similar English terms by the cosine distance among all ICD-9-CM concepts, and the results are showed in Table 6. Most letter-based languages normalize correctly to the term “*dorsalgia*” and character-based language (Chinese, Japanese, and Korean) did not perform as well. Letter-based languages benefited from sharing similar sub-words like “*do-*lor”. The UMLS does not include Chinese terms, but CODER could normalize “背痛” because Chinese shares characters with Japanese. Unfortunately, CODER<sub>ALL</sub> could not handle the unique alphabets of Japanese and Korean well.

**5.3.2 Knowledge-injected term normalization.** To show knowledge can enhance the ability of CODER, we compare the most similar terms extracted by CODER<sub>ENG</sub> and CODER<sub>ENG</sub>- $S_{ij}^{rel}$ .

We list top similar diagnosis terms for “*pufferfish poisoning*” and “*angina*” from ICD-9-CM in Table 7. With additional knowledge from the UMLS, CODER<sub>ENG</sub> can normalize “*pufferfish poisoning*” to “*toxic eff fish/shellfish*”. Both CODER<sub>ENG</sub> and CODER<sub>ENG</sub>- $S_{ij}^{rel}$  can normalize “*angina*” to “*angina pectoris*”. CODER<sub>ENG</sub> further

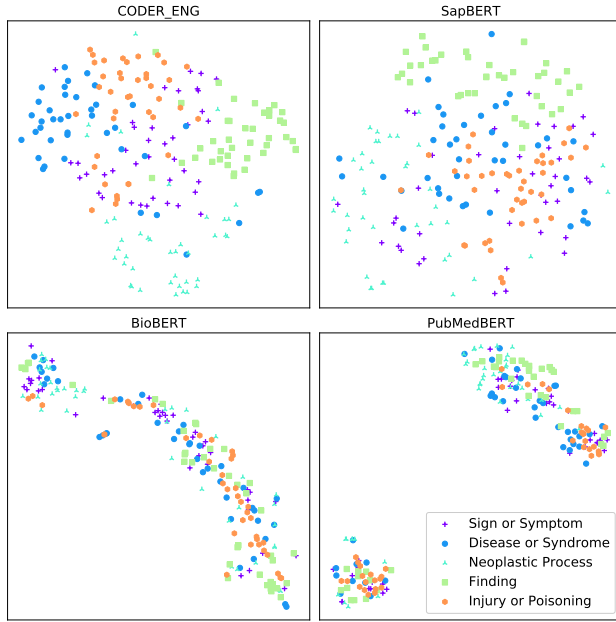


Figure 3: The t-SNE Visualization of different embeddings for concepts from different semantic types.

Table 6: The closest term of C0004604(dorsalgia) in different languages.

Language	Term	Target
Spanish	dolor de espalda	dorsalgia
Hungarian	Hátfájás	dorsalgia
Japanese	背部痛	dorsalgia
Japanese	はいぶつう	puberty
Czech	bolesti zad	dorsalgia
Germany	Rückenschmerzen	dorsalgia
Italian	Dolore dorsale	dorsalgia
Korean	배통	benign neoplasm of ureter
Finnish	Selkäkipu	dorsalgia
French	Dorsalgie	dorsalgia
Dutch	Pijn, rug-	dorsalgia
Norwegian	Ryggsmerter	dorsalgia
Croatian	BOL U LEDIMA	dorsalgia
Swedish	Ryggsmärta	dorsalgia
Chinese	背痛	dorsalgia

finds the “chest pain” is similar to “angina”, which is informative since “chest pain” may be caused by “angina”.

## 6 CONCLUSIONS

In this paper, we present a medical term embedding model CODER. The KG contrastive learning framework learns both term-term similarity and term-relation-term similarity. Evaluations show that CODER substantially outperforms existing medical embeddings in zero-shot term normalization and that it also exhibits good performance in

Table 7: Top 5 similar terms searched from ICD-9-CM.

Top-k	pufferfish poisoning	angina
CODER <sub>ENG</sub>		
1	<b>toxic eff fish/shellfish</b>	<b>angina pectoris</b>
2	poisoning by saluretics	<b>chest pain</b>
3	toxic effect of fusel oil	prinzmetal angina
4	poisoning by paraldehyde	vincent’s angina
5	poisoning by succinimides	angina decubitus
CODER <sub>ENG</sub> -S <sub>rel</sub> <sup>ij</sup>		
1	poisoning by saluretics	<b>angina pectoris</b>
2	poison antivaricose drug	vincent’s angina
3	poisoning by paraldehyde	prinzmetal angina
4	anphylct reaction fish	angina decubitus
5	louping ill	pois coronary vasodilat

semantic similarity and relation classification. CODER<sub>ALL</sub> is the first cross-lingual medical term embedding model, which can help low-resource languages in medical NLP tasks.

## REFERENCES

- [1] Zubair Afzal, Saber A Akhondi, Herman HHBM van Haagen, Erik M Van Mulligen, and Jan A Kors. 2016. Concept recognition in French biomedical text using automatic translation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 162–173.
- [2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 72–78. <https://doi.org/10.18653/v1/W19-1909>
- [3] Andrew L Beam, Benjamin Kompa, Allen Schmalz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. 2020. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing*, Vol. 25. World Scientific, 295–306.
- [4] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. <https://doi.org/10.18653/v1/D19-1371>
- [5] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl\_1 (2004), D267–D270.
- [6] Xiangrui Cai, Jinyang Gao, Kee Yuan Ngiam, Beng Chin Ooi, Ying Zhang, and Xiaojie Yuan. 2018. Medical Concept Embedding with Time-Aware Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 3984–3990. <https://doi.org/10.24963/ijcai.2018/554>
- [7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. (2020).
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 1597–1607. <http://proceedings.mlr.press/v119/chen20j.html>
- [9] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to Train good Word Embeddings for Biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Berlin, Germany, 166–174. <https://doi.org/10.18653/v1/W16-2922>
- [10] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*



- Mining, San Francisco, CA, USA, August 13-17, 2016, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1495–1504. <https://doi.org/10.1145/2939672.2939823>
- [11] Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016 (2016), 41.
  - [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
  - [13] Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics* 121 (2006), 279.
  - [14] Shobeir Fakhraei, Joel Mathew, and José Luis Ambite. 2019. Nseen: Neural semantic embedding for entity normalization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 665–680.
  - [15] Hongchao Fang and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766* (2020).
  - [16] John M Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659* (2020).
  - [17] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779* (2020).
  - [18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 9726–9735. <https://doi.org/10.1109/CVPR42600.2020.00975>
  - [19] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
  - [20] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).
  - [21] Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing Biomedical Embeddings from Language Models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Minneapolis, USA, 82–89. <https://doi.org/10.18653/v1/W19-2011>
  - [22] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics* 55 (2015), 73–81.
  - [23] Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association* 22, 5 (2015), 948–956.
  - [24] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access* (2020).
  - [25] Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29, 22 (2013), 2909–2917.
  - [26] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
  - [27] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9119–9130. <https://doi.org/10.18653/v1/2020.emnlp-main.733>
  - [28] Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. CNN-based ranking for biomedical entity normalization. *BMC bioinformatics* 18, 11 (2017), 79–86.
  - [29] Nut Limsopatham and Nigel Collier. 2016. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1014–1023. <https://doi.org/10.18653/v1/P16-1096>
  - [30] Chin Lin, Yu-Sheng Lou, Dung-Jang Tsai, Chia-Cheng Lee, Chia-Jung Hsu, Ding-Chung Wu, Mei-Chuen Wang, and Wen-Hui Fang. 2019. Projection word embedding model with hybrid sampling training for classifying ICD-10-CM codes: longitudinal observational study. *JMIR medical informatics* 7, 3 (2019), e14499.
  - [31] Yucong Lin, Keming Lu, Yulin Chen, Chuan Hong, and Sheng Yu. 2020. High-throughput relation extraction algorithm development associating knowledge articles and electronic health records. *arXiv preprint arXiv:2009.03506* (2020).
  - [32] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment Pre-training for Biomedical Entity Representations. *arXiv preprint arXiv:2010.11784* (2020).
  - [33] Hanxiao Liu, Yuxin Wu, and Yiming Yang. 2017. Analogical Inference for Multi-relational Embeddings. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2168–2178. <http://proceedings.mlr.press/v70/liu17d.html>
  - [34] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. <https://openreview.net/forum?id=Bkg6RiCqY7>
  - [35] Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep Neural Models for Medical Concept Normalization in User-Generated Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Florence, Italy, 393–399. <https://doi.org/10.18653/v1/P19-2055>
  - [36] Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.), 3111–3119. <https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html>
  - [37] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* 6, 1 (2016), 1–10.
  - [38] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proceedings of LBM* (2013), 39–44.
  - [39] Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2019. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters* 49, 3 (2019), 1239–1256.
  - [40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
  - [41] Nikhil Pattisapu, Sangameshwar Patil, Girish Palshikar, and Vasudeva Varma. 2020. Medical concept normalization by encoding target knowledge. In *Machine Learning for Health Workshop*. PMLR, 246–259.
  - [42] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 58–65. <https://doi.org/10.18653/v1/W19-5006>
  - [43] Naiara Perez, Pablo Accuosto, Àlex Bravo, Montse Cuadros, Eva Martínez-García, Horacio Saggion, and German Rigau. 2020. Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English. *Bioinformatics* 36, 6 (2020), 1872–1880.
  - [44] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. <https://doi.org/10.18653/v1/N18-1202>
  - [45] Minh C. Phan, Aixin Sun, and Yi Tay. 2019. Robust Representation Learning of Biomedical Names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3275–3285. <https://doi.org/10.18653/v1/P19-1317>
  - [46] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
  - [47] Roland Roller, Madeleine Kittner, Dirk Weissenborn, and Ulf Leser. 2018. Cross-lingual candidate search for biomedical concept normalization. *arXiv preprint arXiv:1805.01646* (2018).
  - [48] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. 2018. Dual Attention Matching Network for Context-Aware Feature Sequence Based Person Re-Identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 5363–5372. <https://doi.org/10.1109/CVPR.2018.00562>
  - [49] Muijen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical Entity Representations with Synonym Marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3641–3650.

- <https://doi.org/10.18653/v1/2020.acl-main.335>
- [50] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics* 84 (2018), 93–102.
- [51] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [52] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. 2014. Medical Semantic Similarity with a Neural Language Model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, Jianzhong Li, Xiaoyang Sean Wang, Minos N. Garofalakis, Ian Soboroff, Torsten Suel, and Min Wang (Eds.). ACM, 1819–1822. <https://doi.org/10.1145/2661829.2661974>
- [53] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 5022–5030. <https://doi.org/10.1109/CVPR.2019.00516>
- [54] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1412.6575>
- [55] Sheng Yu, Yumeng Ma, Jessica Grönsbell, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, et al. 2018. Enabling phenotypic big data with PheNorm. *Journal of the American Medical Informatics Association* 25, 1 (2018), 54–60.
- [56] Xiao Zhang, Dejing Dou, and Ji Wu. 2020. Learning Conceptual-Contextual Embeddings for Medical Text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9579–9586.
- [57] Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D Shah, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, et al. 2019. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in brief* 24 (2019), 103838.

## A APPENDICES

### A.1 A summary of compared medical embeddings

We list a summary of medical embeddings in Table 8. Several versions of BioBERT, ClinicalBERT, SciBERT, BlueBERT, and PubMedBERT exist with different training corpora.

We choose BioBERT-base v1.1<sup>2</sup>, Bio\_ClinicalBERT<sup>3</sup>, scibert-scivocab-uncased<sup>4</sup>, BlueBERT-base-uncased-PumMed+MIMIC III<sup>5</sup>, BiomedNLP-PubMedBERT-base-uncased-abstract<sup>6</sup> to compare.

**Table 8: Summary of compared medical embeddings.**

Embedding	# Words	Dim.	Category
cui2vec <sup>7</sup>	108477	500	concept
DeVine-etal-200 <sup>8</sup>	52102	200	concept
claims-codes-hs-300 <sup>9</sup>	14852	300	concept
GoogleNews-vectors <sup>10</sup>	3.0M	300	word
Wiki-pubmed-PMC <sup>11</sup>	5.5M	200	word
BioNLP-2016-win-2 <sup>12</sup>	2.2M	200	word
BioNLP-2016-win-30 <sup>13</sup>	2.2M	200	word
BERT-base-cased <sup>14</sup>	-	768	contextual
BioBERT	-	768	contextual
ClinicalBERT(Huang) <sup>15</sup>	-	768	contextual
ClinicalBERT	-	768	contextual
SciBERT	-	768	contextual
BlueBERT	-	768	contextual
PubMedBERT	-	768	contextual
SapBERT <sup>16</sup>	-	768	contextual

### A.2 English terms normalization results for using [CLS] representation

Results of Cadec and PsyTar for contextual embeddings using [CLS] representation are shown in Table 9. Compared to Table 1, we find using average pooling representation is beneficial for term normalization.

<sup>2</sup><https://drive.google.com/file/d/1R84voFKHfWV9xjzeLzWBmY1uOMYpnyD>

<sup>3</sup>[https://huggingface.co/emilyalsentzer/Bio\\_ClinicalBERT](https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT)

<sup>4</sup>[https://s3-us-west-2.amazonaws.com/ai2-s2-research/scibert/huggingface\\_pytorch/scibert\\_scivocab\\_](https://s3-us-west-2.amazonaws.com/ai2-s2-research/scibert/huggingface_pytorch/scibert_scivocab_)

<sup>5</sup>[https://huggingface.co/bionlp/bluebert\\_pubmed\\_mimic\\_uncased\\_L-12\\_H-768\\_A-12](https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12)

<sup>6</sup><https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract>

<sup>7</sup><https://figshare.com/s/00d69861786cd0156d81>

<sup>8</sup>[https://github.com/clinicalml/embeddings/blob/master/DeVine\\_etal\\_200.txt.gz](https://github.com/clinicalml/embeddings/blob/master/DeVine_etal_200.txt.gz)

<sup>9</sup>[https://github.com/clinicalml/embeddings/blob/master/claims\\_codes\\_hs\\_300.txt.gz](https://github.com/clinicalml/embeddings/blob/master/claims_codes_hs_300.txt.gz)

<sup>10</sup><https://drive.google.com/file/d/0B7XkCwpI5KDYNINUTTISS21pQmM>

<sup>11</sup><http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin>

<sup>12</sup><https://drive.google.com/file/d/0BzMCqpcgEJgiUWs0ZnU0NlFTam8>

<sup>13</sup><https://drive.google.com/file/d/0BzMCqpcgEJgiUWs0ZnU0NlFTam8>

<sup>14</sup>[https://storage.googleapis.com/bert\\_models/2018\\_10\\_18/cased\\_L-12\\_H-768\\_A-12.zip](https://storage.googleapis.com/bert_models/2018_10_18/cased_L-12_H-768_A-12.zip)

<sup>15</sup><https://drive.google.com/file/d/1t8L9w-r88Q5-sfC993x2Tjt1pu--A900>

<sup>16</sup><https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext>

**Table 9: Acc@k for contextual embeddings in Cadec and PsyTar datasets. Contextual embeddings report results with [CLS] representation.**

Embedding	Cadec		PsyTar	
	acc@1	acc@3	acc@1	acc@3
BERT-base-cased	8.03	11.03	7.07	9.66
BioBERT	8.33	11.58	4.36	6.38
ClinicalBERT(Huang)	9.91	14.78	9.02	12.60
ClinicalBERT	7.18	8.84	5.92	7.83
SciBERT	10.61	14.66	8.79	12.55
BlueBERT	11.42	15.13	9.08	12.10
PubMedBERT	11.32	14.79	8.93	13.32
SapBERT	49.26	62.61	44.22	58.62
CODER <sub>ENG</sub>	<b>58.53</b>	<b>75.96</b>	<b>55.27</b>	<b>73.31</b>
CODER <sub>ALL</sub>	52.93	66.53	38.36	52.48

### A.3 MCSM results for using [CLS] representation

We list MCSM scores for contextual embeddings using [CLS] representation in Table 10. Compared to Table 3, average pooling representation achieves better performance on semantic similar measurement except for CODER.

**Table 10: MCSM results for contextual embeddings using [CLS] representation.**

Embedding	PS	DS	NP	CD	FD	IP
BERT-base-cased	8.1	7.2	6.2	9.0	6.3	8.2
BioBERT	8.6	8.3	8.5	9.2	6.9	9.0
ClinicalBERT(Huang)	7.5	6.4	5.7	8.6	5.5	8.5
ClinicalBERT	7.0	7.1	7.2	8.9	5.9	8.0
SciBERT	7.6	6.9	6.2	8.5	6.3	7.5
BlueBERT	7.0	6.3	6.4	8.4	5.9	7.7
PubMedBERT	9.3	9.4	9.1	10.0	8.4	10.0
SapBERT	9.4	8.8	9.3	9.7	9.1	9.8
CODER <sub>ENG</sub>	10.3	<b>9.5</b>	9.6	<b>10.7</b>	9.5	<b>10.1</b>
CODER <sub>ALL</sub>	<b>10.4</b>	<b>9.5</b>	<b>9.8</b>	<b>10.7</b>	<b>9.7</b>	<b>10.1</b>

### A.4 Training parameters of DDBRC

For feature-based, the learning rate is set at 1e-3 for all kinds of embeddings. For fine-tuning, the learning rate is set at 1e-3 for word embeddings and concept embeddings and 2e-5 for contextual embeddings. Batch-size is 512 for feature-based contextual embeddings, word embedding, and concept embeddings. For fine-tuned contextual embeddings, batch-size is 96. Contextual embeddings use **e** to represent terms. All models are trained for 50 epochs.

### A.5 Ablation parameters

In the ablation study, all models are trained with 100K training steps and initialized from PubMedBERT.

We adopt different parameters of MS-loss since the TransE-like term-relation-term similarity function  $S_{ij}^{rel,T}$  is unbounded. The parameter  $\alpha, \beta, \lambda, \epsilon$  of MS-loss are still set to 2, 50, 0.5, 0.1 respectively. We notice that the L2-norm of term-relation-term similarity is close to 28 when the model is initialized, and we let  $\lambda^{rel} = 28$  to avoid exponential function explode. We also set  $\alpha^{rel}, \beta^{rel}, \epsilon^{rel}$  to 0.2, 5, 0.1 for numerical stability. The weight  $\mu$  is set to 0.1 to balance term-term loss and term-relation-term loss.

All other parameters are the same with Section 4.2.