Target Concept Guided Medical Concept Normalization in Noisy User-Generated Texts

Katikapalli Subramanyam Kalyan
Department of Computer Applications
NIT Trichy, India
kalyan.ks@yahoo.com

Sivanesan Sangeetha
Department of Computer Applications
NIT Trichy, India
sangeetha@nitt.edu

Abstract

Medical concept normalization (MCN) i.e., mapping of colloquial medical phrases to standard concepts is an essential step in analysis of medical social media text. The main drawback in existing state-of-the-art approach (Kalyan and Sangeetha, 2020b) is learning target concept vector representations from scratch which requires more training instances. Our model is based on RoBERTa and target concept embeddings. In our model, we integrate a) target concept information in the form of target concept vectors generated by encoding target concept descriptions using SRoBERTa, state-of-the-art RoBERTa based sentence embedding model and b) domain lexicon knowledge by enriching target concept vectors with synonym relationship knowledge using retrofitting algorithm. It is the first attempt in MCN to exploit both target concept information as well as domain lexicon knowledge in the form of retrofitted target concept vectors. Our model outperforms all the existing models with an accuracy improvement up to 1.36% on three standard datasets. Further, our model when trained only on mapping lexicon synonyms achieves up to 4.87% improvement in accuracy.

1 Introduction

Medical concept normalization (MCN) involves learning a model which can assign medical concept from a standard lexicon for the given health related mention. Table 1 shows few examples of concept mentions and corresponding standard concepts from SNOMED-CT lexicon. Normalizing medical concepts finds application in tasks like questions answering, pharmacovigilance, knowledge graph construction etc. In this work, we deal with medical concept normalization in noisy user-generated texts like tweets and online discussion forum posts. With the rising popularity of social media platforms, common public are using these platforms to share information. For example, in twitter people share their health experiences and in websites like AskAPatient.com, public post reviews for the drugs they consume. This valuable health information available in social media platforms can be exploited in applications like pharmacovigilance, public health monitoring etc (Kalyan and Sangeetha, 2020c). In general, most of the common public express their health related concerns in an informal way using colloquial language. For example, *'dizziness'* is expressed as *'head spinning a little'* and *'diarrhoea'* is expressed as *'bathroom with runs'* (Limsopatham and Collier, 2016; Lee et al., 2017). As social media text is highly noisy with irregular grammar and colloquial words, medical concept normalization in social media text is more challenging.

Concept Mention	Standard Concept
lowering of energy	lack of energy (SNOMED ID: 248274002)
felt weak	asthenia (SNOMED ID: 13791008)
very severe pain in arms	pain in upper limb (SNOMED ID: 102556003)
only wanted to sleep	hypersomnia (SNOMED ID: 77692006)

Table 1: Examples of concept mentions and corresponding standard concepts from Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) lexicon.

1.1 Motivation

Most of the existing work in medical concept normalization in social media text ignore valuable target concept knowledge (Limsopatham and Collier, 2016; Lee et al., 2017; Han et al., 2017; Belousov et al., 2017). Recently researchers (Tutubalina et al., 2018; Miftahutdinov and Tutubalina, 2019; Pattisapu et al., 2020; Kalyan and Sangeetha, 2020b) focused on exploiting target concept knowledge in normalizing concepts. The drawbacks in these recent works in integrating target concept

Proceedings of Deep Learning Inside Out (DeeLIO):
The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 64–73
Online, November 19, 2020. ©2020 Association for Computational Linguistics

knowledge in deep learning based medical concept normalization systems are

- Tutubalina et al. (2018) and Miftahutdinov and Tutubalina (2019) exploit target concept knowledge in the form of cosine similarity between tf-idf based vector representations of concept mentions in social media text and concept descriptions from UMLS. However, tf-idf based cosine similarity features between concept mentions and concept descriptions are not effective as concept mentions are noisy, descriptive and colloquial in nature while concept descriptions are expressed in formal language.

- Pattisapu et al. (2020) choose appropriate target concept based on cosine similarity between concept mention and graph embedding based target concept vectors. Here concept mentions are encoded using RoBERTa and then transformed to target concepts embedding space using two fully connected layers. However, a) the quality of graph embedding based target concept vectors depends on the comprehensiveness of mapping lexicon which limits the application of this approach (e.g., MedDRA is less comprehensive compared to SNOMED-CT (Bodenreider, 2009)) b) graph embedding methods used by Pattisapu et al. (2020) generate target concept vectors based on network structure only and completely ignore other information like concept text description and c) when mapping lexicon used is different across datasets, it requires more time and resources to generate target concept vectors using graph embedding methods for each dataset (Kalyan and Sangeetha, 2020b).

- Kalyan and Sangeetha (2020b) learn the vector representations of concept mentions and concepts jointly. The authors randomly assign values to target concept vectors and update them at the time of training. However, learning concept vectors from scratch requires more number of training instances. With less number of training instances, this approach results in poor performance which we illustrate in Section 6.1. This is the current state-of-the-art approach in medical concept normalization in social media text.

Our proposed model overcomes the drawbacks in existing work in utilizing target concept knowledge and answers the following two research questions.

- RQ1 - How to effectively integrate target concept knowledge in deep learning based medical concept normalization system?

- RQ2 - How to utilize domain lexicon knowledge in medical concept normalization?

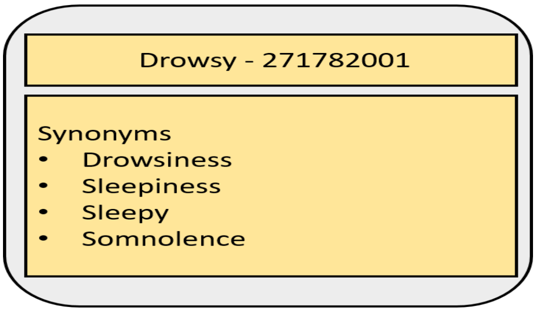

Figure 1: SNOMED-CT Concept and its synonyms. Here, '*Drowsy*' is concept description and '*271782001*' is concept-id.

As shown in Figure 1, every concept has concept-id, description and set of synonyms. To address RQ1, we represent each target concept using fixed length dense vector which is generated by encoding target concept description using SRoBERTa. SRoBERTa (Reimers and Gurevych, 2019) is Siamese network based Sentence RoBERTa model trained on NLI+Multi NLI and STS datasets. It is state-of-the-art sentence embedding model which encodes sequence of words into dense fixed length vectors in a way that sequences which are in close meaning are also close in embedding pace . To address RQ2, we retrofit target concept vectors produced by SRoBERTa using synonyms from mapping lexicon. Retrofitting algorithm (Faruqui et al., 2015) enriches concept vectors with synonym relationship knowledge from domain lexicon.

In our model we encode a) input concept mentions using RoBERTa and b) target concepts using SRoBERTa and enrich them with synonym relationship knowledge. We compute similarity vector in which each value is equal to cosine similarity between vectors of concept mentions and all the target concepts. Finally, the cosine similarity values are normalized and the target concept with maximum similarity is chosen. During training, the vectors of target concepts are not updated. We evaluate our model on three standard MCN datasets CADEC, PsyTAR and SMM4H2017 and achieve accuracy improvements up to 1.36%. Further, our model when trained only using mapping lexicon synonyms achieves up to 4.87% improvement in accuracy. The key aspects of our work are

- A simple approach to integrate both target concept information and domain lexicon knowledge in medical concept normalization in the form of retrofitted target concept vectors.

- Our model achieves state-of-the-art performance on three standard medical concept normalization datasets.

- Our model when trained using mapping lexicon synonyms only, achieves up to 4.87% improvement in accuracy which shows that our approach to generate target concept vectors is better than graph embedding based approach (Pattisapu et al., 2020) or learning from scratch (Kalyan and Sangeetha, 2020b).

2 Related Work

2.1 Medical Concept Normalization

Traditional concept normalization systems used string matching (Aronson, 2001; McCallum et al., 2005; Tsuruoka et al., 2007) or machine learning approaches (Leaman et al., 2013; Leaman and Lu, 2014). These methods perform poorly in case of instances with no words in common between concept mention and concept description. With the introduction of embedding models like Word2vec (Mikolov et al., 2013) and ELMo (Peters et al., 2018) which can encode syntactic and semantic information, researchers focused on exploiting embeddings in normalizing medical concepts. For example, Limsopatham and Collier (2016) used CNN and RNN models with word2vec embeddings. Subramanyam and Sangeetha (2020) proposed a model based on BiLSTM and clinical ELMo embeddings. Han et al. (2017) used hierarchical character LSTM on the top of character embeddings while Belousov et al. (2017) used Multinomial Logistic Regression classifier on the top of embeddings inferred from various corpora.

In recent times, unsupervised pre-trained models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) achieved significant improvements in most of the natural language processing tasks. Most of the recent work in medical concept normalization (Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020a; Pattisapu et al., 2020; Kalyan and Sangeetha, 2020b) in social media text is based on BERT and RoBERTa. Miftahutdinov and Tutubalina (2019) experimented with BERT and cosine similarity based semantic features, Kalyan and Sangeetha (2020a) experimented with various general and domain specific BERT models combined with highway network layer. Pattisapu et al. (2020) normalize medical concepts using RoBERTa and graph embedding based concept vectors while approach of Kalyan and Sangeetha (2020b) involves learning the vectors representations of target concepts along with input concept mentions. Our approach is similar to (Kalyan and Sangeetha, 2020b) by choosing target concept which has maximum cosine similarity with the input concept mention. However unlike Kalyan and Sangeetha (2020b) method which learns target concept vectors from scratch, we use retrofitted target concept vectors which are generated using SRoBERTa and then enriched with synonym relationship knowledge from domain lexicon. It is the first work to exploit both target concept information and domain lexicon knowledge effectively in MCN in the form of retrofitted target concept vectors.

2.2 Sentence Embeddings

Sentence embeddings encode sequence of words into dense fixed size vector. Some of the popular approaches are averaging word vectors, encoder-decoder based skip thought (Kiros et al., 2015), InferSent (Conneau et al., 2017) which is Siamese BiLSTM+max pooling trained on SNLI, transformer based Universal Sentence Encoder (Cer et al., 2018). Recently, Reimers and Gurevych (2019) proposed SRoBERTa, Siamese network based Sentence RoBERTa model and it is trained on NLI + MultiNLI datasets followed by STS dataset. It is a state-of-the-art sentence embedding model which encodes sequence of words into dense fixed length vector in a way that sequences which are close in meaning are also close in embedding pace.

2.3 Retrofitting algorithm

Vector representations generated by neural embedding models are rich in syntactic and semantic information but lack valuable relationship knowledge from semantic lexicons. To enrich vector representations with relationship knowledge, Faruqui et al. (2015) proposed retrofitting algorithm. It is simply a post-processing step and can be applied to vectors generated using any embedding model. It learns retrofitted concept vectors $\{v_1, v_2, v_3, .., v_n\}$ from concept vectors $\{\hat{v_1}, \hat{v_2}, \hat{v_3}, ..., \hat{v_n}\}$ by iteratively minimizing distance between (i) retrofitted vector v_i and its counterpart $\hat{v}_i$ and (ii) retrofitted vector v_i and all its neighbors v_j. The objective function is

$$\sum_{i=1}^{n} \left[\alpha_i \left\| v_i - \hat{v}_i \right\|^2 + \sum_{(i,j) \in E} \beta_{ij} \left\| v_i - v_j \right\|^2 \right] \tag{1}$$

Here retrofitted vectors v_i are initialized with values of concept vectors $\hat{v}_i$ and then updated iteratively by minimizing the objective function.

3 Datasets

Our proposed model is evaluated on three standard MCN datasets of noisy user-generated texts. Out of these, CADEC (Karimi et al., 2015) and PsyTAR (Zolnoori et al., 2019) datasets contain concept mentions gathered from user-generated AskAPatient.com reviews and SMM4H2017 (Sarker et al., 2018) contains adverse drug reaction (ADR) mentions extracted from twitter.

CADEC: Karimi et.al released CSIRO Adverse Drug Event Corpus (CADEC) having user posted drug reviews gathered from AskAPatient (Karimi et al., 2015). The annotators manually identified concept mentions and mapped them to SNOMED-CT concepts which resulted in a corpus of 6754 concept mentions and 1029 SNOMED-CT codes. As 66% of instances are common in train and test splits in the random folds of this dataset released by Limsopatham and Collier (2016), Tutubalina et al. (2018) split this dataset into five folds[1] with no overlap.

PsyTAR: Zolnoori et al. (2019) released PsyTAR corpus which includes 887 user generated psychiatric drug reviews collected from AskAPatient. This dataset includes manually identified 6556 concept phrases which are mapped to 618 concepts in SNOMED-CT. Zolnoori et al. (2019) released random folds of this dataset. However, 56% of instances are common in train and test in these folds. So, Miftahutdinov and Tutubalina (2019) create custom folds of this dataset[2] to reduce the overlap between train and test sets.

SMM4H017: Sarker et al. (2018) released this dataset[3] of ADR mentions for subtask3 of SMM4H2017 shared task organized by Health Language Processing Lab @ University of Pennsilvaniya. Initially, twets containing generic and trade names of drugs were collected. Then, ADR mentions were manually identified and mapped to MedDRA concepts. In this corpus, train set consists of 6500 ADR phrases and 472 unique MedDRA codes, test set consists of 2500 ADR phrases and 254 MedDRA codes.

The significant overlap between train and test sets in random folds of CADEC and PsyTAR datasets can result in bias and contribute to high performance of model (Lee et al., 2017; Kalyan and Sangeetha, 2020a). So, we evaluate our approach on custom folds of PsyTAR and CADEC datasets in addition to SMM4H2017 dataset, like the recent previous works (Pattisapu et al., 2020; Kalyan and Sangeetha, 2020b)

4 Methodology

4.1 Model Description

Our model is based on RoBERTa and target concept embeddings. Initially we compute vector representations of input phrase and concepts in standard lexicon using RoBERTa and SRoBERTa respectively. We further enrich target concept vectors with synonym relationship from domain lexicon using retrofitting algorithm. Then, we find cosine similarity between vectors of concept mention and all the target concepts. Finally, the concept mention is mapped to concept with maximum similarity. Figure 2 gives an overview of our proposed model.

Target Concept Representation

We use SRoBERTa, state-of-the-art sentence embedding model to compute target concept representations and then inject synonym relationship using retrofitting algorithm to get target concept vector $\boldsymbol{e_c} \in \mathbb{R}^h$.

$$\boldsymbol{e_c} = Retrofit(SRoBERTa(concept)) \tag{2}$$

Concept Mention Representation

Learning quality representation of concept mentions is a key step in medical concept normalization. We use RoBERTa, which is an improved version of BERT with large training batch sizes and more

[1] https://cutt.ly/Gi6kka6

[2] https://doi.org/10.5281/zenodo.3236318

[3] https://data.mendeley.com/datasets/rxwfb3tysd/2

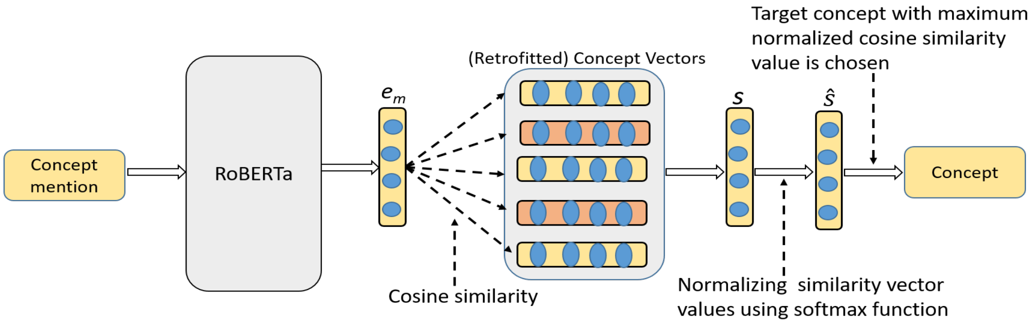
Concept mention
RoBERTa
e_m
(Retrofitted) Concept Vectors
Cosine similarity
s
$\hat{s}$
Target concept with maximum normalized cosine similarity value is chosen
Normalizing similarity vector values using softmax function
Concept

Figure 2: Overview of our proposed model for medical concept normalization in noisy user-generated texts. $\boldsymbol{e_m}$ - RoBERTa encoded input concept mention, $\boldsymbol{s}$ - similarity vector computed based on cosine similarity between the vectors of input phrase and concepts in standard lexicon, $\hat{\boldsymbol{s}}$ - normalized cosine similarity vector.

training corpus, to compute input concept mention representation $\boldsymbol{e_m} \in \mathbb{R}^h$.

$$\boldsymbol{e_m} = RoBERTa(mention) \tag{3}$$

We find similarity vector based on cosine similarity between vectors of input pharse and concepts in standard lexicon. Finally, we normalize all the cosine similarity values using softmax which result in normalized similarity vector $\hat{\boldsymbol{s}} \in \mathbb{R}^C$.

$$\hat{\boldsymbol{s}} = [\hat{\boldsymbol{s}}_{\boldsymbol{i}}]_{i=1}^{C} \tag{4}$$

Here C represents total number of unique target concepts in the dataset, $\hat{\boldsymbol{s}}_{\boldsymbol{i}} = Softmax(f(\boldsymbol{e_m}, \boldsymbol{e_{ci}}))$ where the function f() represents cosine similarity and $\boldsymbol{e_{ci}}$ represents vector of the concept c_i. We train the model using AdamW optimizer (Loshchilov and Hutter, 2019) which minimize cross entropy loss (L_{CE}) between normalized similarity vector $\hat{\boldsymbol{s}}$ and the ground truth vector $\boldsymbol{s}$. During training, we freeze the vectors of target concepts.

$$L_{CE} = -\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{C} s_j^i log(\hat{s}_j^i) \tag{5}$$

4.2 Implementation Details

We do basic pre-processing steps like lower-casing, removing non-ASCII and special characters in concept mention and concept descriptions. We remove unnecessary words like 'nos', 'unspecified' and 'finding' in concept descriptions. In case of concept mentions, we do additional pre-processing steps like removing repeating characters (e.g., sooo much → so much) , replacing medical acronyms[4] ('ra' → 'rheumatoid arthritis') and contractions (isn't → is not) with full forms.

Pattisapu et al. (2020) treat synonyms in mapping lexicon as concept mention and augment the training set with the labeled instances generated from synonyms. However, we augment training set with synonyms of less frequently occurring concepts only. In case of CADEC and PsyTAR datasets, we use synonyms from the mapping lexicon SNOMED-CT. In case of SMM4H2017 dataset, we use synonyms from UMLS Metathesaurus as synonyms are very few in number in MedDRA. For each concept in MedDRA, we find the corresponding concept unique identifier(CUI) in UMLS and then gather all the associated synonyms excluding non-English synonyms.

In case of retrofitting algorithm, we choose number of iterations = 10 as suggested by the authors. Further, we use the implementation[5] provided by the authors. As there is no official validation set in case of all the three datasets, we use 10% of the augmented training set for validation. We find optimal hyperparameter values by performing random search over the range of hyperparameter values. During training, we freeze target concept vectors. We implement all our models in PyTorch using transformers package from huggingface (Wolf et al., 2019).

4.3 Evaluation Metrics

In case of all the three datasets, standard evaluation metric is accuracy (Miftahutdinov and Tutubalina, 2019; Pattisapu et al., 2020; Kalyan and Sangeetha, 2020b). In case of CADEC and PsyTAR datasets which are multi-fold, reported accuracy is average accuracy across all five folds.

[4] Gathered from UMLS Methathesaurus, Wikipedia and https://www.acronymslist.com/cat/medical-acronyms.html

[5] https://github.com/mfaruqui/retrofitting

4.4 Comparison with existing methods

Here, we compare our approach with the following existing methods.

Hierarchical Character-LSTM Han et al. (2017) use hierarchical character level LSTM to normalize the concept mentions. Intially, they generate character level word representations using LSTM over embeddings of characters and their classes and then apply bidirectional LSTM over these word representations to generate contextual word vectors. Finally, vector obtained by max-pooling of contextual word vectors is given to fully connected softmax layer.

Multinomial LR Belousov et al. (2017) generate concept mention vector representation as average of three weighted vectors of words in the concept mention. Here, word weights are based on inverse document frequencies of words and word vectors are obtained as average of GoogleNews, twitter and drugtwitter embeddings. With these mention representations as input, Multinomial Logistic Regression classifier assigns the concepts.

BERT + Cosine Semantic Features Miftahutdinov and Tutubalina (2019) generate representation of concept mention using BERT. To integrate target concept knowledge , the authors generate semantic features based on cosine similarity between tf-idf vector representations of concept mention and all the target concepts in the dataset. Finally, the output of BERT and cosine semantic features are concatenated and given to fully connected softmax layer which assigns the concepts.

BERT + Highway Network Layer Kalyan and Sangeetha (2020a) experiment with various general and domain specific BERT models for medical concept normalization. The output of BERT model is passed through highway network layer to eliminate the unnecessary information and then passed through fully connected softmax layer to get the target concept.

RoBERTa + Graph based Concept Vectors Pattisapu et al. (2020) generate target concept vectors using graph embedding algorithms. They train RoBERTa based model which embeds input concept mention into the embedding space of target concept vectors. For a given input phrase, the nearest standard concept in embedding space is assigned.

RoBERTa + Random Concept Vectors Kalyan and Sangeetha (2020b) propose a model based on RoBERTa which jointly learns the representations of concept mention and the standard concepts. The authors randomly initialize the target concept vectors and then they are updated during training. The standard concept with maximum cosine similarity with input phrase is chosen.

4.5 Models

RoBERTa We generate the representations of input concept mention using RoBERTa. We experiment with both variants of RoBERTa namely RoBERTa-base and RoBERTa-large. In both the cases, the size of concept mention vector is equal to the hidden vector size i.e., 768 in case of RoBERTa-base and 1024 in case of RoBERTa-large.

+ Concept Vectors (CV) We generate target concept vectors by encoding their descriptions using SRoBERTa. In case of a) RoBERTa-base model, we use target concept vectors generated by ‘roberta-base-nli-stsb-mean-tokens’ and b) RoBERTa-large model, we use target concepts generated by ‘roberta-large-nli-stsb-mean-tokens’.

+ Retrofitted Concept Vectors(RCV) We enrich target concepts generated by SRoBERTa with synonym relationship knowledge from mapping lexicon using retrofitting algorithm.

5 Results

Our proposed model is evaluated on the standard MCN datasets CADEC, PsyTAR and SMM4H2017. The performance of our model and existing models is presented in Table 2. From Table 2, we notice that our proposed model achieves the best results of 86.40%, 85.04% and 91.73% across CADEC, PsyTAR and SMM4H2017 datasets. Our model outperforms existing methods with accuracy improvement up to 1.36%. The existing state-of-the-art model Kalyan and Sangeetha (2020b) learns target concept vectors from scratch and so it requires more number of training instances. Our model outperforms the approach of Kalyan and Sangeetha (2020b) (i) up to 1.9% in case of base version and (ii) up to 1.36% in case of large version. The use of retrofitted concept vectors improved performance only in case of SMM4H2017. The performance of retrofitted concept vectors depends on the number of available synonyms for each concept.

Method	CADEC	PsyTAR	SMM4H2017
Existing Methods			
(Han et al., 2017)	-	-	87.20
(Belousov et al., 2017)	-	-	87.70
(Miftahutdinov and Tutubalina, 2019)	79.83	77.52	89.28
(Kalyan and Sangeetha, 2020a)	82.62	-	-
(Pattisapu et al., 2020)	83.18	82.42	-
(Kalyan and Sangeetha, 2020b)$^{\pi}$	82.60	81.90	90.15
(Kalyan and Sangeetha, 2020b)$^{\Pi}$	85.49	83.68	90.84
Our Method			
RoBERTa-base+ CV$^{\gamma}$	84.53	82.41	91.34
RoBERTa-base+ RCV$^{\delta}$	84.11	82.34	91.19
RoBERTa-large + CV$^{\gamma}$	**86.40**	**85.04**	91.19
RoBERTa-large+ RCV$^{\delta}$	86.04	85.02	**91.73**

Table 2: Performance of our mdoel and existing methods on CADEC, PsyTAR and SMM4H2017 datasets. π - model based on Roberta-base and Π - model based on Roberta-large. γ - concept vectors generated using SRoBERTa and δ - concept vectors generated using SRoBERTa and then retrofitted using synonym relationship from domain lexicon.

The synonyms for SMM4H2017 are gathered from UMLS Metathesaurus and as they are more number in number compared to SNOMED-CT synonyms, retrofitted concept vectors improve accuracy only in case of SMM4H2017 (Roberta-large). In future, we would like to see whether using UMLS synonyms instead of SNOMED-CT synonyms improve performance in case of CADEC and PsyTAR datasets also.

6 Analysis and Discussion

6.1 Training only on mapping lexicon synonyms

There will be a set of synonyms for each concept in mapping lexicon. Table 3 shows some of the concepts and corresponding synonyms from SNOMED-CT lexicon. We consider each synonym as user-generated concept mention and generate labeled instances from mapping lexicon synonyms.

To show the performance of our model in the absence of human annotated instances in training set, we train our model using labeled instances generated from mapping lexicon synonyms and then evaluate our model on the corresponding test set. Tabel 4 shows the performance of our model and existing models across three datasets. As reported in the table, our model outperforms existing methods with accuracy improvement up to 4.46% and 4.87% across CADEC and PsyTAR datasets respectively.

From Table 4, we infer that among the three approaches, Kalyan and Sangeetha (2020b) achieved the lowest performance in case of CADEC and PsyTAR datasets. When compared to Kalyan and Sangeetha (2020b), the performance of a) Pattisapu et al. (2020) is 9.42% and 9.87% higher b) our approach is 13.88% and 14.74% higher across CADEC and PsyTAR datasets respectively. Kalyan and Sangeetha (2020b) learn the vector representations of concept mentions and concepts jointly. The authors randomly assigned values to target concepts and then updated them during training. However, learning concept vectors from scratch requires more number of training instances. As the number of training instances generated from synonyms is less in number, this approach results in poor performance.

In case of SMM4H2017, Kalyan and Sangeetha (2020b) achieved the best performance of 63.28% which is 2.55% more than our approach. Here as the number of training instances generated from synonyms is more in number, Kalyan and Sangeetha (2020b) outperformed our approach. This shows that learning target concept vectors from scratch is effective only when training instances are more in number.

6.2 Failure Analysis

Here we analyse the reasons for the wrong predictions given by our best performing model. For this, we check all the failure cases in CADEC dataset.

Our model failed to handle the concept mentions which are misspelled words of ground truth con-

Concept-ID	Concept Description	Concept Synonyms
60119000	Exhaustion	Washed out, Worn out
278040002	Loss of hair	Thinning hair, Falling hair
386705008	Lightheadedness	Feels light headed, Dizziness light headed , Lightheaded
131148009	Bleeding	Haemorrhage, Hemorrhage
247640008	Unable to think clearly	Muddled thought , Muddled thinking
102897001	Feeling intoxicated	Feeling drunk, Feeling groggy

Table 3: Concepts and their synonyms from SNOMED-CT lexicon

Method	CADEC	PsyTAR	SMM4H2017
Existing Methods			
(Pattisapu et al., 2020)	64.80	58.4	-
(Kalyan and Sangeetha, 2020b)$^{\pi}$	51.55	45.77	55.75
(Kalyan and Sangeetha, 2020b)$^{\Pi}$	55.38	48.53	**63.28**
Ours			
Roberta-base + CV$^{\gamma}$	62.44	59.47	58.78
Roberta-base + RCV$^{\delta}$	63.73	60.14	57.31
Roberta-large + CV$^{\gamma}$	**69.26**	63.06	58.82
Roberta-large + RCV$^{\delta}$	69.14	**63.27**	60.73

Table 4: Performance of our model and existing methods when trained only on mapping lexicon synonyms. π - model based on Roberta-base and Π - model based on Roberta-large. γ - concept vectors generated using SRoBERTa. δ - concept vectors generated using SRoBERTa and then retrofitted using synonym relationship from domain lexicon.

cepts. For example, the concept mention *'insomina* is mapped to *'nausea - 422587007 '* instead of the ground truth concept *'insomnia - 193462001'*. Similarly, the concept mentions *'naseua'*, *'fatique'*, *'insommnia'*, *'diziness'*, *'nausia'* and *'diarreah'* are not mapped to the ground truth concepts *'nausea - 422587007'*, *'fatigue - 84229001 '*, *'insomnia - 193462001 '*, *'dizziness - 404640003'*, *'nausea - 422587007'* and *'diarrhea - 62315008'* respectively. Here, all the concept mentions are misspelled words of the ground truth concepts.

In some of the cases, our model assigned concepts which are more specific than the ground truth concepts. For example, our model mapped the concepts mentions *'pain so bad'*, *'so much pain'*, *'worse pain'* and *'pain bad'* to the concept *'severe pain - 76948002'* rather than the ground truth *'pain - 22253000'*. Here we observe that in case of all these concept mentions, the concept *'severe pain'* is more specific and hence appropriate compared to the ground truth *'pain'*.

In few cases, our model assigned concepts which are closely related to the ground truth concept. For example, the concept mention *'difficult to concentrate* is assigned to the concept *'unable to concentrate - 60032008'* instead of the ground truth concept *'poor concentration - 26329005'*. Here the predicted and ground truth concepts are closely related. Similarly, *'could not walk across the room'* is assigned to *'unable to walk - 282145008'* instead of *'walking disability - 228158008'*.

One more case in which our model failed is when the concept mention is an abbreviation of the ground truth concept. For example, the concept mention *'ibu'* is assigned to the concept *'ubidecarenone'* and the ground truth concept is *'ibuprofen*. Here, *'ibu'* is an abbreviation of *'ibuprofen'*.

6.3 Limitations

In case of SMM4H2017 dataset, we find the corresponding CUI for each MedDRA concept and include all the associated synonyms excluding non-English synonyms. Here the limitation is that, some CUIs can be mapped to more than one MedDRA concept. For example, *'C0020649 (Hypotension)'* can be mapped to both the MedDRA concepts *'10021097 (Hypotension)'* and *'10005734 (Blood pressure decreased)'*. Similarly, *'C0036974 (Shock)'* can be mapped to both the MedDRA concepts *'10009192 (Circulatory collapse)'* and

'10034567 (Peripheral circulatory failure)'.

7 Conclusion

Here, we propose a model based on RoBERTa and target concept embeddings to normalize concepts in medical related user-generated texts. Our model integrates target concept knowledge as well domain lexicon knowledge in a simple and novel way. The existing state-of-the-art approach (Kalyan and Sangeetha, 2020b) exploits target concept knowledge by learning vector representations of target concepts from scratch. As target concept vectors are learned from scratch, this approach requires more training instances and it performs poorly with less number of training instances. Our model exploits target concept information and domain lexicon knowledge in the form of retrofitted target concept vectors. We encode target concepts using SRoBERTa and enrich these concept vectors with synonym relationship knowledge from standard lexicon using retrofitting algorithm. Our model outperforms all the existing methods and achieves significant improvements on three standard datasets.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Maksim Belousov, William Dixon, and Goran Nenadic. 2017. Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task. In *SMM4H@ AMIA*, pages 54–58.

Olivier Bodenreider. 2009. Using snomed ct in combination with meddra for reporting signal detection and adverse drug reactions reporting. In *AMIA Annual Symposium Proceedings*, volume 2009, page 45. American Medical Informatics Association.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

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)*, pages 4171–4186.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Sifei Han, Tung Tran, Anthony Rios, and Ramakanth Kavuluru. 2017. Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter. In *SMM4H@ AMIA*, pages 49–53.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020a. Bertmcn: Mapping colloquial phrases to standard medical concepts using bert and highway network. Technical report, EasyChair.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020b. Medical concept normalization in user generated texts by learning target concept embeddings. *arXiv preprint arXiv:2006.04014*.

Katikapalli Subramanyam Kalyan and S Sangeetha. 2020c. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman and Zhiyong Lu. 2014. Automated disease normalization with low rank approximations. In *Proceedings of BioNLP 2014*, pages 24–28.

Kathy Lee, Sadid A Hasan, Oladimeji Farri, Alok Choudhary, and Ankit Agrawal. 2017. Medical concept normalization for online user-generated texts. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 462–469. IEEE.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrew McCallum, Kedar Bellare, and Fernando Pereira. 2005. A conditional random field for discriminatively-trained finite-state string edit distance. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 388–395.

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*, pages 393–399.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Nikhil Pattisapu, Sangameshwar Patil, Girish Palshikar, and Vasudeva Varma. 2020. Medical Concept Normalization by Encoding Target Knowledge. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 246–259. PMLR.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

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)*, pages 3973–3983.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 2018. Data and systems for medication-related text classification and concept normalization from twitter: insights from the social media mining for health (smm4h)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283.

Kalyan Katikapalli Subramanyam and S Sangeetha. 2020. Deep contextualized medical concept normalization in social media text. *Procedia Computer Science*, 171:1353 – 1362. Third International Conference on Computing and Network Communications (CoCoNet'19).

Yoshimasa Tsuruoka, John McNaught, Jun'i; chi Tsujii, and Sophia Ananiadou. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768–2774.

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*, 84:93–102.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, Mike Conway, et al. 2019. A systematic approach for developing a corpus of patient reported adverse drug events: a case study for ssri and snri medications. *Journal of biomedical informatics*, 90:103091.