Medical Concept Normalization in User-Generated Texts by Learning Target Concept Embeddings

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 helps in discovering standard concepts in free-form text i.e., maps health-related mentions to standard concepts in a clinical knowledge base. It is much beyond simple string matching and requires a deep semantic understanding of concept mentions. Recent research approach concept normalization as either text classification or text similarity. The main drawback in existing a) text classification approach is ignoring valuable target concepts information in learning input concept mention representation b) text similarity approach is the need to separately generate target concept embeddings which is time and resource consuming. Our proposed model overcomes these drawbacks by jointly learning the representations of input concept mention and target concepts. First, we learn input concept mention representation using RoBERTa. Second, we find cosine similarity between embeddings of input concept mention and all the target concepts. Here, embeddings of target concepts are randomly initialized and then updated during training. Finally, the target concept with maximum cosine similarity is assigned to the input concept mention. Our model surpasses all the existing methods across three standard datasets by improving accuracy up to 2.31%.

1 Background

Internet users use social media to voice their views and opinions. Medical social media is a part of social media in which the focus is limited to health and related issues (Pattisapu et al., 2017). User generated texts in medical social media include tweets, blog posts, reviews on drugs, health related question and answers in discussion forums. This rich source of data can be utilized in many health related applications to enhance the quality of services provided (Kalyan and Sangeetha, 2020b).

Medical concept normalization aims at discovering standard medical concepts in free-form text. In this task, health related mentions are mapped to standard concepts in a clinical knowledge base. For example, the concept mention *'hard to stay awake'* is mapped to the standard concept *'drowsy'*. The common public express their health related conditions in an informal way using layman terms while clinical knowledge base contains concepts expressed in scientific language. This variation (colloquial vs scientific) in the languages of common public and knowledge bases makes concept normalization an essential step in understanding user-generated texts. This task is much beyond simple string matching as the same concept can be expressed in a descriptive way using colloquial words or in multiple ways using aliases, acronyms, partial names and morphological variants. Further, noisy nature of user-generated texts and the short length of health-related mentions make the task of concept normalization more challenging.

Research in medical concept normalization started with string matching techniques (Aronson, 2001; McCallum et al., 2005; Tsuruoka et al., 2007) followed by machine learning techniques (Leaman et al., 2013; Leaman and Lu, 2014). The inability of these methods to consider semantics into account shifted research towards deep learning methods with embeddings as input (Limsopatham and Collier, 2016; Lee et al., 2017; Tutubalina et al., 2018; Subramanyam and Sangeetha, 2020). For example, Lee et al. (2017) and Tutubalina et al. (2018) experimented with RNN on the top of domain specific embeddings. Further, lack of large labeled datasets and necessity to train deep learning models like CNN or RNN from scratch (except embeddings) shifted research towards using pretrained language models like BERT and RoBERTa (Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020a; Pattisapu et al., 2020). Miftahut-

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 18–23
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

dinov and Tutubalina (2019) experimented with BERT based fine-tuned models while Kalyan and Sangeetha (2020a) provided a comprehensive evaluation of BERT based general and domain specific models. The approach of Pattisapu et al. (2020) is based on RoBERTa (Liu et al., 2019) and graph embedding based target concept vectors. The main drawbacks in existing work are :

- text classification approach (Limsopatham and Collier, 2016; Lee et al., 2017; Subramanyam and Sangeetha, 2020; Kalyan and Sangeetha, 2020a) is not exploiting target concepts information in learning input concept mention representation . However, recent work in various natural language processing and computer vision tasks highlights the importance of exploiting target label information in learning input representation. (Rodriguez-Serrano et al., 2013; Akata et al., 2015; Wang et al., 2018; Pappas and Henderson, 2019; Liu et al., 2020).

- text similarity approach of Pattisapu et al. (2020) is the need to generate target concept embeddings separately using graph embedding methods. This is time and resource consuming when different vocabularies are used for mapping in different data sets (e.g., SNOMED-CT is used in CADEC (Karimi et al., 2015) and PsyTAR (Zolnoori et al., 2019) datasets, MedDRA (Mozzicato, 2009) is used in SMM4H2017 (Sarker et al., 2018)). Moreover, the quality of generated concept embeddings using graph embedding methods depends on the comprehensiveness of vocabulary. For example, MedDRA is less fine grained compared to SNOMED-CT (Bodenreider, 2009). This requirement of comprehensive vocabulary limits the effectiveness of this approach.

Our model normalizes input concept mention by jointly learning the representations of input concept mention and target concepts. By learning the representations of target concepts along with input concept mention, our model a) exploits target concepts information unlike existing text classification approaches (Tutubalina et al., 2018; Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020a) and b) eliminates the time and resource consuming process of separately generating target concept embeddings unlike existing text similarity approach (Pattisapu et al., 2020). Our key contributions are :

- We propose a simple and novel approach which exploits the target concepts information in normalizing concept mention by jointly learning the representations of input concept mention and all the target concepts. It is the first work in medical concept normalization which jointly learns the representations of input concept mention and the target concepts.

- Our model achieves the best results across three standard data sets surpassing all the existing methods with an accuracy improvement of up to 2.31%.

2 Methodology

2.1 Model Description

Our model normalizes concept mentions in two phases. First, it learns input concept mention representation using RoBERTa (Liu et al., 2019). Second, it finds cosine similarity between embeddings of input concept mention and all the target concepts. Here, embeddings of target concepts are randomly initialized and then updated during training. Finally, the target concept with maximum cosine similarity is assigned to the input concept mention.

Input concept mention is encoded into a fixed size vector $\boldsymbol{m} \in \mathbb{R}^d$ using RoBERTa. RoBERTa is a contextualized embedding model pre-trained on 160 GB of text corpus. It consists of an embedding layer followed by a sequence of transformer encoders (Liu et al., 2019).

$$\boldsymbol{m} = RoBERTa(mention) \tag{1}$$

Input concept mention vector $\boldsymbol{m}$ is transformed into cosine similarity vector $\boldsymbol{q} \in \mathbb{R}^N$ by finding cosine similarity between m and randomly initialized embeddings $\{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3}, \ldots \boldsymbol{c_N}\}$ of all target concepts $\{\boldsymbol{C_1}, \boldsymbol{C_2}, \ldots \boldsymbol{C_N}\}$ where $\boldsymbol{c_i} \in \mathbb{R}^d$ and N represents total number of unique target concepts in the dataset. During training, the target concept embeddings and parameters of RoBERTa are updated. Here d is equal to size of hidden state vector in RoBERTa (768 in RoBERTa-base and 1024 in RoBERTa-large).

$$\boldsymbol{q} = [q_i]_{i=1}^{N} \, where \, q_i = CS(\boldsymbol{m}, \boldsymbol{c_i}) \tag{2}$$

Here $i = 1, 2, 3, \ldots N$ and the function $CS()$ represents cosine similarity defined as

$$CS(\boldsymbol{m}, \boldsymbol{c}) = \frac{\sum_{i=1}^{d} m_i \times c_i}{\sqrt{\sum_{i=1}^{d} (m_i)^2} \times \sqrt{\sum_{i=1}^{d} (c_i)^2}} \tag{3}$$

Cosine similarity vector $\boldsymbol{q}$ is normalized to $\hat{\boldsymbol{q}}$ using softmax function.

$$\hat{\boldsymbol{q}} = Softmax(\boldsymbol{q}) \tag{4}$$

Finally, model parameters and target concept embeddings are updated using AdamW optimizer (Loshchilov and Hutter, 2019) which minimizes cross entropy loss ($\mathfrak{L}$) between normalized cosine similarity vector $\hat{\boldsymbol{q}}$ and one hot encoded ground truth vector $\boldsymbol{p} \in \mathbb{R}^N$. Here M represents number of training instances.

$$\mathfrak{L} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} p_j^i log(\hat{q}_j^i) \tag{5}$$

2.2 Evaluation Metric

We evaluate our normalization system using accuracy metric, as in the previous works (Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020a; Pattisapu et al., 2020). Accuracy represents the percentage of correctly normalized mentions. In case of CADEC (Karimi et al., 2015) and PsyTAR (Zolnoori et al., 2019) datasets which are multi-fold, reported accuracy is average accuracy across folds.

3 Experimental Setup

3.1 Implementation Details

Pre-processing of input concept mentions include a) removal of non-ASCII and special characters b) normalizing words with more than two consecutive repeating characters (e.g., sleeep → sleep) and c) replacing English contraction and medical acronym words with corresponding full forms (e.g., can't → cannot, bp → blood pressure). The list of medical acronyms is gathered from acronymslist.com and Wikipedia. Pattisapu et al. (2020) generate additional labeled instances by considering synonyms in mapping lexicon as user-geneated concept mentions and augment training set with these labeled instances. However, we don't augment the training set with any additional labeled instances generated from mapping lexicon and we use only the training instances available in the datasets . We choose 10% of training set for validation and find optimal hyperparameter values using random search. We use AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 3e-5. The final results reported are based on the optimal hyperparameter settings. To implement our model, we choose PyTorch framework and transformers library (Wolf et al., 2019).

3.2 Datasets

SMM4H2017 : This dataset is released for task3 of SMM4H 2017 (Sarker et al., 2018) shared tasks. It consists of ADR phrases extracted from twitter using drug names as keywords and then mapped to Preferred Terms (PTs) from MedDRA. In this, training set includes 6650 phrases assigned with 472 PTs and test set includes 2500 phrases assigned with 254 PTs.

CADEC: CSIRO Adverse Drug Event Corpus (CADEC) includes user generated medical reviews related to Diclofenac and Lipitor (Karimi et al., 2015). The manually identified health related mentions are mapped to target concepts in SNOMED-CT vocabulary. The dataset includes 6,754 mentions mapped to one of the 1029 SNOMED-CT codes. As the random folds of CADEC dataset created by Limsopatham and Collier (2016) have significant overlap between train and test instances, Tutubalina et al. (2018) create custom folds [1] of this dataset with minimum overlap.

PsyTAR: Psychiatric Treatment Adverse Reactions (PsyTAR) corpus includes psychiatric drug reviews obtained from AskaPatient (Zolnoori et al., 2019). Zolnoori et al. (2019) manually identify 6556 health related mentions and map them to one of 618 SNOMED-CT codes. Due to significant overlap between train and test sets of random folds released by Zolnoori et al. (2019), Miftahutdinov and Tutubalina (2019) create custom folds[2] of this dataset with minimum overlap.

We evaluate our model using SMM4H2017, custom folds of CADEC and PsyTAR datasets.

4 Results

Table 1 provides a comparison of our model and the existing methods across three standard concept normalization datasets CADEC, PsyTAR and

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

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

Method	CADEC	PsyTAR	SMM4H17
(Tutubalina et al., 2018)	70.05	-	-
(Subramanyam and Sangeetha, 2020)	75.12	-	-
(Han et al., 2017)	-	-	87.20
(Belousov et al., 2017)	-	-	87.70
(Miftahutdinov and Tutubalina, 2019)	79.83	77.52	89.64
(Kalyan and Sangeetha, 2020a)	82.62	-	-
(Pattisapu et al., 2020)	83.18	82.42	-
Roberta-base + concept embeddings$^{\perp}$	82.60	81.90	90.15
Roberta-large + concept embeddings$^{\perp}$	**85.49** (**2.31** ↑)	**83.68** (**1.26** ↑)	**90.84** (**1.2** ↑)

Table 1: Accuracy of existing methods and our proposed model across CADEC, PsyTAR and SMM4H2017 datasets. $\perp$ - concept embeddings are randomly initialized and then updated during training.

SMM4H2017. The first seven rows represent existing systems and the next two rows represent our approach. Our model achieves new state-of-the-art accuracy of 85.49%, 83.68% and 90.84% across three datasets. Our model outperforms the existing state-of-the-art method of Pattisapu et al. (2020) with accuracy improvement of 2.31%, 1.26% and 1.2% respectively. We didn't augment the training set with labeled instances generated out of synonyms from mapping lexicon like Pattisapu et al. (2020), but still our approach achieved significant improvements. State-of-the-art results achieved by our model across three standard datasets illustrate that learning target concept representations along with input mention representations is simple and much effective compared to separately generating target concept representations using graph embedding methods and then using them.

5 Analysis

Here, we discuss merits and demerits of our proposed method.

5.1 Merit Analysis

We illustrate the effectiveness of our approach in the following two cases.

- In case I, existing methods map the concept mention '*no concentration*' to a closely related target concept '*Poor concentration (26329005)*' instead of the correct target concept '*Unable to concentrate (60032008)*'. Similarly, '*sleepy*' is mapped to '*hypersomnia (77692006)*' instead of '*drowsy (271782001)*'.

- In case II, '*horrible pain*' is mapped to abstract target concept '*Pain (22253000)*' instead of fine-grained target concept '*Severe pain (76948002)*'. Similarly, '*fatigue in arms*' is mapped to '*fatigue (84229001)*' instead of '*muscle fatigue (80449002)*'.

In both the cases, existing methods are unable to exploit target concept information effectively and fail to assign the correct concept. However, our approach exploits target concept information by jointly learning representations of input concept mention and target concepts and hence assigns the concepts correctly.

5.2 Demerit Analysis

Our model aims to map health related mentions to standard concepts. We observe the predictions of our model and identify the following errors.

- In case I, errors are related to insufficient number of training instances. For example, '*hard to stay awake*' is assigned with more frequent concept '*insomnia (193462001)*' instead of the ground truth concept '*drowsy (271782001)*'. Similarly '*muscle cramps in lower legs*' is assigned with '*cramp in lower limb (449917004)*' instead of '*cramp in lower leg (449918009)*'.

- In case II, errors are related to the inability in learning appropriate representations for domain specific rare words. For example, the mentions '*pruritus*' and '*hematuria*' are assigned to completely unrelated concepts '*Tinnitus (60862001)*' and '*diarrhea (62315008)*' respectively.

6 Conclusion

In this work, we deal with medical concept normalization in user generated texts. Our model overcomes the drawbacks in existing text classification and text similarity approaches by jointly learning the representations of input concept mention and target concepts. By learning target concept representations along with input concept mention representations, our approach a) exploits valuable target concepts information unlike existing text classification approaches and b) eliminates the need to separately generate target concept embeddings unlike existing text similarity approach. Our model surpasses all the existing methods across three standard datasets by improving accuracy up to 2.31%. In future, we would like to explore other possible options to include target concept information which may further improve the results.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2015. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438.

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.

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.

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. 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.

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.

Xien Liu, Song Wang, Xiao Zhang, Xinxin You, Ji Wu, and Dejing Dou. 2020. Label-guided learning for text classification. *arXiv preprint arXiv:2002.10772*.

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.

Patricia Mozzicato. 2009. Meddra. *Pharmaceutical Medicine*, 23(2):65–75.

Nikolaos Pappas and James Henderson. 2019. Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.

Nikhil Pattisapu, Manish Gupta, Ponnurangam Kumaraguru, and Vasudeva Varma. 2017. Medical persona classification in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 377–384.

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.

Jose A Rodriguez-Serrano, Florent Perronnin, and France Meylan. 2013. Label embedding for text recognition. In *BMVC*, pages 5–1.

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.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331.

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.