Social Media Medical Concept Normalization using RoBERTa in Ontology Enriched Text Similarity Framework

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

Pattisapu et al. (2020) formulate medical concept normalization (MCN) as text similarity problem and propose a model based on RoBERTa and graph embedding based target concept vectors. However, graph embedding techniques ignore valuable information available in the ontology like concept description and synonyms. In this work, we enhance the model of Pattisapu et al. (2020) with two novel changes. First, we use retrofitted target concept vectors instead of graph embedding based vectors. It is the first work to leverage both concept description and synonyms to represent concepts in the form of retrofitted target concept vectors in text similarity framework based social media MCN. Second, we generate both concept and concept mention vectors with same size which eliminates the need of dense layers to project concept mention vectors into the target concept embedding space. Our model outperforms existing methods with improvements up to 3.75% on two standard datasets. Further when trained only on ontology synonyms, our model outperforms existing methods with improvements up to 14.61%. We attribute these improvements to the two novel changes introduced.

1 Introduction

With the rise of internet and easy accessibility, social media has become primary choice to share information. Social media includes generic platforms like twitter, facebook and health related platforms like *AskAPatient.com* and *Patient.info*. Most of the common public express their health related issues in a descriptive way using informal language. For example, a person suffering from diarrhoea expresses it as *"bathroom with runs"*. Some of the colloquial health mentions along with standard concepts is given in Table 1. However all the knowledge in clinical ontologies is available in standard medical terms. Due to this variation in the style of languages used, it is necessary to map health related mentions expressed in colloquial language to corresponding concepts in standard clinical ontology. This mapping of colloquial mentions to standard concepts is referred to as medical concept normalization (MCN) and is useful in applications like identification of adverse drug reactions, clinical paraphrasing, question answering and public health monitoring (Lee et al., 2017; Pattisapu et al., 2020). However, normalizing user-generated health related mentions is difficult due to the colloquial language and noisy nature.

Research in medical concept normalization in social media text started with phrase-based machine translation model of Limsopatham and Collier (2015). Previously, most of the existing work approach medical concept normalization in social media text as supervised multi-class text classification (Limsopatham and Collier, 2016; Tutubalina et al., 2018; Subramanyam and Sivanesan, 2020; Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020). In this approach, initially concept mention representation vector is learned using any of the deep learning models and then it is given to fully connected softmax layer to get the predicted concept. Some of the models are based on shallow neural networks like CNN or RNN with word embeddings as input (Limsopatham and Collier, 2016; Tutubalina et al., 2018; Subramanyam and Sivanesan, 2020) and the rest of

Concept Mention	Standard Concept
feel so off	Depersonalization (ID: 79499004)
rapid heart beat	Tachycardia (ID: 3424008)
unable to relax	Restlessness (ID:162221009)

Table 1: Colloquial health mentions and their standard concepts in SNOMED-CT (Donnelly, 2006).

Proceedings of the First Knowledgeable Natural Language Processing Workshop, pages 21–26

the models are based on BERT (Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020) . For example, Limsopatham and Collier (2016) proposed models based on CNN and RNN with out-of-domain word embeddings as input, Tutubalina et al. (2018) experimented with attention based RNN model on the top of in-domain word embeddings and Subramanyam and Sivanesan (2020) experimented with bidirectional RNNs and ELMo embeddings. Miftahutdinov and Tutubalina (2019) experimented with BERT and tf-idf based semantic features while Kalyan and Sangeetha (2020) proposed model based on BERT (Devlin et al., 2019) and Highway Networks (Srivastava et al., 2015).

The main drawbacks in classification based MCN systems are a) completely ignoring target concept information by representing target concepts as meaningless one hot vectors. b) with the addition of new concepts every year, the number of concepts in clinical knowledge base is increasing. To accommodate new concepts, these models have to be re-trained from scratch which is time-taking and computationally expensive process (Pattisapu et al., 2020). To overcome the drawbacks in multi-class classification framework to normalize medical concepts, Pattisapu et al. (2020) formulate MCN as a text similarity problem and propose a model based on RoBERTa (Liu et al., 2019) and graph embedding based target concept vectors. Initially, all the target concept vectors are generated using graph embedding techniques. Then they fine-tune a RoBERTa based model which learns concept mention vector and then projects it into target concepts embedding space using two dense fully connected layers. Finally, the closest target concept to the concept mention in the embedding space is chosen.

The main drawbacks in the model of Pattisapu et al. (2020) are

- Graph embedding techniques leverage only the network structure and completely ignore other valuable information associated with concepts like concept description and synonyms. Moreover it is time and resource consuming process to generate target concept vectors using graph embedding techniques.

- As vectors of concept mentions and concepts are generated with different sizes, it is necessary to project vectors of concept mentions into the embedding space of target concept using dense layers to find the nearest target concept to the given concept mention. As parameters of these dense layers are randomly initialized, good number of training instances are required to learn these dense layers parameters. With limited number of training instances, these parameters are not learned well which limits the performance of model as illustrated in Section 4.

Like Pattisapu et al. (2020), we approach MCN as text similarity problem. The contribution of this paper is the two novel changes we introduce in the original model of Pattisapu et al. (2020) to overcome the drawbacks and further improve the performance.

- First, we use retrofitted embeddings to represent concepts. Each concept has description and set of synonyms. We encode concept descriptions using SRoBERTa (Reimers and Gurevych, 2019) and then enhance the generated concept embeddings with the injection of synonym relationship knowledge using retrofitting algorithm (Faruqui et al., 2015) and concept synonyms. Moreover, it is easy and fast to compute retrofitted embeddings. It is the first work to leverage both concept description and synonyms to represent concepts in the form of retrofitted target concept vectors in text similarity framework based social media MCN.

- Second, we generate concept vectors and concept mention vectors with same size which eliminates the need of dense layers for projecting concept mentions vectors into target concept embedding space.

Following Pattisapu et al. (2020), we conduct experiments on two publicly available MCN datasets and achieve improvements up to 3.75%. Further when trained only on mapping lexicon synonyms, our model outperforms existing methods with improvements up to 14.61%. We attribute these improvements to the two novel changes introduced.

2 Proposed Method

Initially, all the target concept vectors are generated using SRoBERTa (Reimers and Gurevych, 2019) and retrofitting algorithm (Faruqui et al., 2015). Then we fine-tune a RoBERTa model which learns representation vector of concept mention. Finally,

Concept-ID	Concept Description	Synonyms
278040002	Loss of hair	falling hair, thinning hair
60862001	Tinnitus	ringing in ears, noise in ears
77692006	Hypersomnia	sleeps too much, excessive sleep hypersomnia, excessive sleepiness
60119000	Exhaustion	washed out, worn out

Table 2: Some of the SNOMED-CT concepts with concept-id (unique medical code), description (fully specified name) and synonyms.

the target concept which is closest (based on cosine similarity) to the concept mention in the embedding space is chosen. We refer to our model as "**O**ntology **E**nriched **T**ext **S**imilarity Framework based **R**oBERTa (OETSR)" .

Each concept has a description and set of synonyms as shown in Table 2. We generate concept vectors in two phases. First, we encode concept descriptions using SRoBERTa . SRoBERTa is a state-of-the-art sentence embedding model which maps concept descriptions to vectors such that related concepts are closer in embedding space. Second, we enchance the quality of target concept vectors with the addition of synonym relationship knowledge using retrofitting algorithm and concept synonyms.

$$\boldsymbol{p} = Retrofit(SRoBERTa(concept), synonyms) \quad (1)$$

Learning concept mention representation is a key step in medical concept normalization (Limsopatham and Collier, 2016; Subramanyam and Sivanesan, 2020; Pattisapu et al., 2020). Like Pattisapu et al. (2020), we use RoBERTa to learn concept mention representations. RoBERTa is a variant of BERT model trained on 160GB text data with better training strategies.

$$\boldsymbol{q} = RoBERTa(mention) \quad (2)$$

We train the model using AdamW optimizer which minimizes cosine embedding loss (L) between concept mention vector, $\boldsymbol{q} \in R^H$ and target concept vector, $\boldsymbol{p} \in R^H$. Here, H is the hidden vector size in RoBERTa. During training, we freeze the target concept vectors.

$$L = 1 - CosineSimilarity(\boldsymbol{p}, \boldsymbol{q}) \quad (3)$$

During inference, we encode concept mention using our fine-tuned RoBERTa model. The concept mention is mapped to the closest target concept (based on cosine similarity) in the embedding space.

3 Experimental Details

3.1 Datasets

CADEC : CSIRO Adverse Drug Event Corpus (CADEC) dataset consists of 6754 colloquial health related mentions gathered from askapatient.com and 1029 unique SNOMED-CT codes (Karimi et al., 2015). The domain experts manually identified all the health related mentions like *'terrible pain in shoulders'* and mapped them to medical codes in SNOMED-CT vocabulary. We evaluate our model on the five fold dataset[1] created from these annotations.

PsyTAR : Zolnoori et al. (2019) gathered psychiatric medicines related reviews from askapatient.com and created this dataset. It consists of 6556 colloquial health related mentions and 618 unique SNOMED-CT codes. Miftahutdinov and Tutubalina (2019) created five fold dataset[2] from these annotations and released it publicly.

SNOMED-CT Synonyms: SNOMED-CT is one of the commonly used medical lexicons which includes around 0.35M concepts. Each medical concept has unique id (code), concept description (fully specified name) and set of synonyms. Each synonym can be treated as a health mention. To show the performance of our model in the absence of manually annotated instances, we train our model on the dataset created from these synonyms and then evaluate on CADEC and PsyTAR datasets. All the results are reported in Table 4.

3.2 Implementation Details

As concept mentions are noisy in nature, we lowercase the text and remove unnecessary special characters and non-ASCII characters. Further, we

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

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

Method	CADEC				PsyTAR			
	Acc@1	Acc@3	Acc@5	Acc@10	Acc@1	Acc@3	Acc@5	Acc@10
Existing Methods								
(Tutubalina et al., 2018)	70.05	-	-	-	-	-	-	-
(Subramanyam and Sivanesan, 2020)	75.12	-	-	-	-	-	-	-
(Miftahutdinov and Tutubalina, 2019)	79.83	-	-	-	77.52	-	-	-
(Kalyan and Sangeetha, 2020)	82.62	-	-	-	-	-	-	-
(Pattisapu et al., 2020)	83.18	-	-	-	82.42	-	-	-
Ours								
OETSR (Roberta-base)$^{\Phi}$	84.37	89.28	91.51	93.43	82.86	89.11	91.11	93.53
OETSR (Roberta-large)$^{\Phi}$	**86.16** (**2.98** ↑)	91.05	93.06	94.94	**85.20** (**2.78** ↑)	91.36	92.83	94.85
OETSR (Roberta-base)$^{\Pi}$	84.96	89.78	92.13	93.86	83.42	89.54	92.09	93.89
OETSR (Roberta-large)$^{\Pi}$	**86.93** (**3.75** ↑)	91.84	93.61	95.19	**85.76** (**3.34** ↑)	92.05	93.27	95.08

Table 3: Comparison of existing methods and our model. Φ - model is trained using training instances + SNOMED-CT synonyms like Pattisapu et al. (2020) and Π model is trained using training instances + UMLS synonyms.

normalize the words with consecutive repeating characters (e.g., feeeel to feel) and replace all the medical acronym words with corresponding full forms. To generate target concept vectors using SRoBERTa, we use sentence-transfomers[3] python library. We use SRoBERTa model trained using NLI (Bowman et al., 2015) + Multi NLI (Williams et al., 2018) datasets followed by further training on STSb (Cer et al., 2017) dataset. We run retrofitting algorithm for ten iterations as suggested by the authors. There is no official validation set for CADEC and PsyTAR datasets. So, we find optimal hyper-parameter values through random search over 10% of training instances as validation set like Pattisapu et al. (2020). We use PyTorch deep learning framework (Paszke et al., 2019) and Transformers library (Wolf et al., 2019) to implement our models.

3.3 Evaluation Metrics

We choose top-k accuracy as evaluation metric following the existing work (Miftahutdinov and Tutubalina, 2019; Kalyan and Sangeetha, 2020; Pattisapu et al., 2020) in medical concept normalization in social media text. Acc@k is 1 if top k predicted concepts include the ground truth concept otherwise 0. We evaluate our model using Acc@1 and Acc@3. As CADEC and PsyTAR datasets are five-fold, reported accuracy is the average of accuracy obtained across the folds.

4 Results

Following (Pattisapu et al., 2020), we conduct experiments on CADEC and PsyTAR datasets. The experimental results are reported in Table 3. As mentioned in Table 3, our model achieves 86.16% and 85.20% on CADEC and PsyTAR datasets respectively. The current state-of-the-art model Pattisapu et al. (2020) achieves 83.18% and 82.42% on CADEC and PsyTAR datasets respectively. Our model outperforms existing methods with improvements of a) 2.98% and 2.78% when trained using training instances + SNOMED-CT synonyms and b) 3.75% and 3.34% when trained using training instances + UMLS synonyms. As the number of labeled instances generated from UMLS synonyms is more compared to the labeled generated from SNOMED-CT synonyms, more improvements are achieved when the model is trained using training instances + UMLS synonyms.

Further, we would like to check how well our proposed model performs when there is no human annotated instances in the training set. For this, we train our model using only labeled instances generated from mapping lexicon synonyms and evaluate on CADEC and PsyTAR datasets. As mentioned in Table 4, our model achieves 69.47% and 65.31% on CADEC and PsyTAR datasets. The current state-of-the-art model Pattisapu et al. (2020) achieves 64.8% and 58.4% on CADEC and PsyTAR datasets respectively. Our model outperforms existing methods with improvements of a) 4.67% and 6.91% when trained using SNOMED-CT synonyms and b) 5.66% and 14.61% when trained using UMLS synonyms. We attribute these improvements to the novel changes introduced by us in the text similarity framework based model of Pattisapu et al. (2020).

[3]https://github.com/UKPLab/sentence-transformers

Method	CADEC				PsyTAR			
	Acc@1	Acc@3	Acc@5	Acc@10	Acc@1	Acc@3	Acc@5	Acc@10
Existing Methods								
(Pattisapu et al., 2020)	64.80	-	-	-	58.40	-	-	-
Ours								
OETSR (Roberta-base)$^{\Phi}$	63.23	75.21	79.96	85.94	60.15	74.33	80.00	86.60
OETSR (Roberta-large)$^{\Phi}$	**69.47** (**4.67** ↑)	81.10	84.76	87.99	**65.31** (**6.91** ↑)	80.95	85.06	88.85
OETSR (Roberta-base)$^{\Pi}$	67.98	78.73	82.49	86.26	66.75	81.17	84.83	89.44
OETSR (Roberta-large)$^{\Pi}$	**70.46** (**5.66** ↑)	83.15	86.13	89.89	**73.01** (**14.61** ↑)	85.08	87.82	90.80

Table 4: Comparison of existing methods and our model. Here model is trained using ontology synonyms and evaluated on the corresponding test sets. Φ - model is trained using SNOMED-CT synonyms like Pattisapu et al. (2020) and Π - model is trained using UMLS synonyms.

5 Analysis and Discussion

5.1 Error Analysis

In this paper, we develop a model which learns to map user-generated concept mentions to standard concepts in clinical knowledge base. To find the reasons for wrong mappings done by our model, we manually check all the erroneous mappings.

Our model failed in cases where the concept mention and predicted concept are exactly the same. For example, the concept mention *'weight gain'* is mapped to 'weight gain' but the ground truth concept is 'excessive weight gain'. This wrong mapping could be due to wrong annotation or interpretation of the mention depends on its context.

In some cases, our model assigned concept which is more specific compared to the ground truth concept. For example, the mentions *'at times felt very anxious'* and *'very anxious'* are mapped to 'severe anxiety' but the ground truth is 'anxiety'. Here the assigned concept is more specific and hence appropriate.

In some cases, our model assigned abstract concept rather than specific concept. For example, our model assigned the mention *'sweat more'* to the concept 'sweating' rather than the ground truth concept 'excessive sweating'. Here the concept 'excessive sweating' is more specific than the concept 'sweating'.

Some of the erroneous mappings occurred when concept mention and the predicted concept overlap. For example, the mention *'anti-constipating'* is mapped to 'constipation' but the ground truth concept is 'diarrhea'.

6 Conclusion

In this work, we come up with a text similarity based model to normalize colloquial health related mentions in user-generated posts. Pattisapu et al. (2020) formulate MCN as text similarity problem and propose a model based on RoBERTa and graph based target concept vectors. Our model is an enhancement of the original model of (Pattisapu et al., 2020) with two simple and novel changes which improve the performance up to 3.75%. We use retrofitted target concept vectors to represent concepts which leverage both concept description and synonyms, unlike graph embedding techniques. Moreover, it is easy and faster to compute retrofitted target concept vectors compared to graph embedding based target concept vectors. In future, we would like to explore options like distant supervision to generate additional training examples.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

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.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

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.

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

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.

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. 2015. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1675–1680.

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*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

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.

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.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Kalyan Katikapalli Subramanyam and Sangeetha Sivanesan. 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).

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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.