Proceedings of Machine Learning Research 116:246–259, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Medical Concept Normalization by Encoding Target Knowledge

Nikhil Pattisapu NIKHIL.PATTISAPU@RESEARCH.IIIT.AC.IN
Language Technologies Research Centre
International Institute of Information Technology, Hyderabad

Sangameshwar Patil SANGAMESHWAR.PATIL@TCS.COM
Tata Research Development and Design Centre, Pune

Girish Palshikar GK.PALSHIKAR@TCS.COM
Tata Research Development and Design Centre, Pune

Vasudeva Varma VV@IIIT.AC.IN
Language Technologies Research Centre
International Institute of Information Technology, Hyderabad

Abstract

Medical concept normalization aims to map a variable length message such as, '*unable to sleep*' to an entry in a target medical lexicon, such as '`Insomnia`'. Current approaches formulate medical concept normalization as a supervised text classification problem. This formulation has several drawbacks. First, creating training data requires manually mapping medical concept mentions to their corresponding entries in a target lexicon. Second, these models fail to map a mention to the target concepts which were not encountered during the training phase. Lastly, these models have to be retrained from scratch whenever new concepts are added to the target lexicon. In this work we propose a method which overcomes these limitations. We first use various text and graph embedding methods to encode medical concepts into an embedding space. We then train a model which transforms concept mentions into vectors in this target embedding space. Finally, we use cosine similarity to find the nearest medical concept to a given input medical concept mention. Our model scales to millions of target concepts and trivially accommodates growing target lexicon size without incurring significant computational cost. Experimental results show that our model outperforms the previous state-of-the-art by 4.2% and 6.3% classification accuracy across two benchmark datasets. We also present a variety of studies to evaluate the robustness of our model under different training conditions.

1. Background

Social media is being increasingly used for patient care, patient support (Attai et al., 2015), pharmacovigilance (Nikfarjam et al., 2015), treatment (Hawn, 2009), enhancement of professional networks (Ventola, 2014), public health monitoring (Paul et al., 2016) and medical education (Cheston et al., 2013). Medical social media is the subset of social media where the interests of the group is dedicated to healthcare. Medical social media spans generic channels such as web logs, chat rooms, Twitter, Facebook, YouTube, and LinkedIn, as well as specific portals which are restricted to healthcare issues such as Mayo Clinic Social Media Network, *patient.info* and *caregiving.com*.

Identifying the medical concepts from medical social media posts is of great value to several organizations. For instance, pharmaceutical firms could use them to identify the adverse events associated with a particular drug (pharmacovigilance). Hospitals and clinics could use them to identify what are patient groups discussing and offer timely advice. Governments could use it to study and improve population health.

The authors of medical social media are usually multi-lingual, multi-cultural and have varying expertise levels and backgrounds. It is therefore difficult for them to adhere to a common standard medical terminology. The variability in medical language, excessive usage of acronyms, jargon, and spell variations allows one to express the same concept in a variety of different ways. These factors make identification and disambiguation of medical concepts a very hard task. Embeddings trained on medical text (Moen and Ananiadou, 2013; Zhu et al., 2018; Alsentzer et al., 2019) have been proved to be useful for such tasks. The task of Medical Concept Normalization (MCN) aims to map a variable length message, such as *'it feels like ma head is bursting'* to a medical concept in a target lexicon, such as `'Headache'`[1].

Limsopatham and Collier (2015) proposed the use of phrase based machine translation model for this task. They trained a model which translates a medical concept mention (informal text) to its corresponding medical term (formal text). Limsopatham and Collier (2016) were amongst the first to set MCN as a supervised text classification task. They have trained multiple deep neural models using pretrained word embeddings along with Convolutional Neural Networks(CNN) and Recurrent Neural Networks (RNN). Tutubalina et al. (2018) proposed a model which uses attention mechanism with Recurrent Neural Networks. The aforementioned models use word-level neural network methods, which are efficient at learning informal expression features but fail to learn character structure features inside words and ignore the Out-of-vocabulary (OOV) words. In order to overcome this problem, Niu et al. (2019) present a multi-task character-level attentional network model for MCN.

Recently, Miftahutdinov and Tutubalina (2019) proposed three models for this task. The first model uses Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM) with an attention mechanism and a hyperbolic tangent activation function on top of pretrained word embeddings to obtain vector representation of concept mentions. The second model uses a bidirectional layer with attention on top of deep contextualized word representations ELMo (Peters et al., 2018) and the third model uses the multilayer bidirectional transformer encoder BERT (Devlin et al., 2018) to extract the vector representation of concept mentions. Finally, they train a softmax based classifier using these representations. They also propose a variant for each of the aforementioned models where they concatenate these representations with semantic similarity features based on prior knowledge from the UMLS Metathesaurus (Bodenreider, 2004). Out of the three models, the BERT based model is the current state-of-the-art for MCN and it achieves the highest classification accuracy of 79.83% and 77.52% across two benchmark datasets.

[1] We use *italics* to denote mentions and `typewriter` to denote medical concept

Previous approaches such as (Limsopatham and Collier, 2016; Tutubalina et al., 2018; Miftahutdinov and Tutubalina, 2019) formulate MCN as a supervised text classification problem. This formulation has several drawbacks. First, creating training data requires manually mapping medical concept mentions to entries in a target lexicon such as SNOMED-CT[2] which is effort intensive. Second, these models fail to map a mention to target concepts which were not encountered during the training phase. The number of medical concepts increase with advancements in human knowledge. Figure 1 depicts the number of unique medical concepts in SNOMED-CT from 2002 to 2019. Current models have to be retrained from scratch whenever new concepts are added to the target lexicon, which is computationally expensive. In this work, we build an MCN model which overcomes these limitations. Our model scales to millions of target concepts and trivially accommodates growing target lexicon size without incurring significant computational cost. Experimental results show that our models outperforms the previous state-of-the-art on two benchmark datasets. We also present a variety of studies to evaluate the robustness of our model under different training conditions. We would release the code and the trained models to facilitate research in this direction.

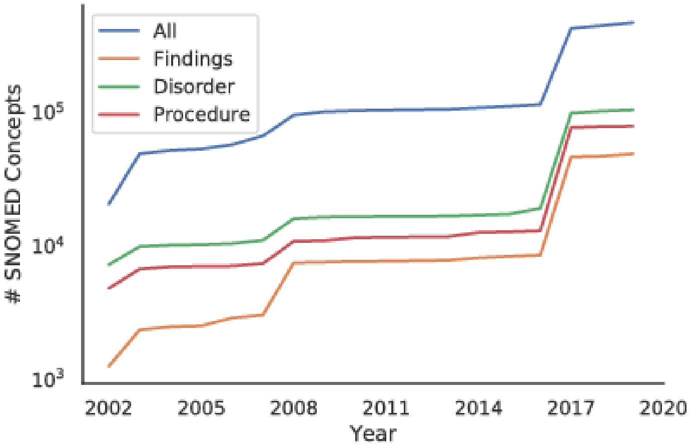

Figure 1: Number of unique SNOMED-CT Concepts across the last 18 years. A base-10 logarithmic scale is used for the Y-axis. Findings, Disorders and Procedures are the major concept types in SNOMED-CT. Color viewing advised.

2. Method

Our method is divided into two stages. In the first stage, we encode all SNOMED-CT medical concepts such as `Headache, SCTID: 25064002` into fixed size embeddings, such that similar concepts are closer in the embedding space. In the second stage, we train

[2] SNOMED-CT is an acronym for Systematized Nomenclature of Medicine – Clinical Terms. The Jan 2019 version of International SNOMED-CT contains more than 450,000 unique medical concepts

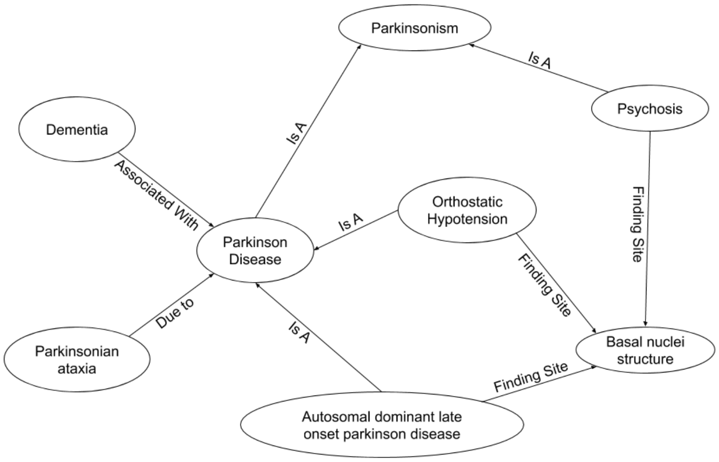

Figure 2: A sample subgraph of the SNOMED-CT Depicting the connected Medical Concepts

a model which transforms a given input concept mention such as *it feels like ma head is bursting* into an embedding in the above mentioned space. Finally, we use cosine similarity to find the nearest medical concept to a given input medical concept mention.

We now describe various methods to encode target concepts into embeddings. Our aim is to map medical concepts to embeddings such that related concepts are closer in the target embedding space. The idea of mapping the target labels to vectors has been previously used to solve text and image classification problems (Akata et al., 2013; Lin et al., 2014; Wang et al., 2018; Akata et al., 2015). However, our work is the first to explore target embeddings for MCN. For the extent of this work, we assume that the target lexicon is SNOMED-CT. SNOMED-CT contains medical concepts indexed by a unique concept ID (such as `42539006`). Each medical concept is associated with a concept description (such as `Parkinson disease`). We work with two distinct views of SNOMED-CT. First, SNOMED-CT as a collection of strings representing the medical concepts and second, SNOMED-CT as a graph where each vertex is a unique medical concept and related concepts are connected through labeled edges. Figure 2 depicts a sample graph view of SNOMED-CT concepts.

Sections 2.1 and 2.2 propose different methods to encode SNOMED-CT concepts by exploiting the lexical and semantic relationships between them.

2.1. Text Embedding Methods

AvgEmb method encodes text by averaging the pre-trained word embeddings of all the words present in it. For the extent of this work, we use the pretrained embeddings released by Mikolov et al. (2013).

BERT uses bidirectional transformer based neural model to solve the task of masked language modeling (Devlin et al., 2018). It generates low dimensional contextualized token embeddings which were shown to be useful for several NLP tasks such as text classification, sentiment analysis, etc. BERT model uses sub-word level information to generate token embeddings, thereby circumventing the problem of OOV words.

Universal Sentence Encoders (USE) use transformer based encoders to encode sentences into embedding vectors (Cer et al., 2018). The encoder uses attention to compute context-aware word embeddings which are then aggregated to obtain the sentence embedding. The sentence embedding is fed to several downstream tasks such as natural language inference and sentence classification. The encoder is trained to foster transfer learning to other NLP tasks.

Embeddings from Language Models (ELMo) uses bi-directional LSTM based encoders to encode a sentence into a fixed size representation (Peters et al., 2018). The model is trained to obtain context aware word embeddings by working on sequence of characters thereby avoiding the problem of OOV words at inference stage.

We obtain embeddings for each SNOMED-CT concept by encoding the corresponding concept description using the pretrained AvgEmb, BERT, USE and ELMo models.

2.2. Graph Embedding Methods

Graph embedding methods embed vertices of a graph such that the similarity in the original graph is approximated by the similarity in the embedding space. The similarities between any two vertices is defined based on whether or not the vertices are connected, share neighbours or have similar structural roles.

DeepWalk (Perozzi et al., 2014) uses random walks to generate sequences of vertices (vertex sentences) which are subsequently fed to a skip-gram model to learn the embeddings corresponding to the vertices.

Node2Vec (Grover and Leskovec, 2016) uses biased random walks to optimize a neighborhood preserving objective function such that the nodes which are highly interconnected and the nodes with similar roles in the graph are closer in the embedding space.

LINE (Tang et al., 2015) tries to directly optimize the vertex embeddings based on one hop and two hop random walk probabilities. Their objective function is designed to preserve first and second order proximity thereby preserving the local and global network structure respectively.

HARP (Chen et al., 2018) proposes a meta-strategy for embedding vertices of a graph such that they preserve the higher-order structural features. They use graph coarsening to create a hierarchy of smaller graphs such that the smaller graphs preserve the global structure of original graph. The vertex embeddings of the coarsest graph are then obtained using DeepWalk, Node2Vec or LINE. They then iteratively prolong and refine the vertex embeddings from the coarsest to the finest graph and finally obtain the vertex embeddings corresponding to the finest (original) graph.

A comprehensive survey of various graph embedding approaches, their performance analysis, time and space complexities are detailed by Goyal and Ferrara (2018). We formulate SNOMED-CT as a directed graph as shown in Figure 2, where vertices are the concepts and relationships between the concepts form the edges. We then train DeepWalk, Node2Vec, LINE and HARP models to obtain the concept embeddings. To the best of our knowledge, our work is the first to explore SNOMED-CT graph embeddings for MCN.

2.3. Transforming Concept Mentions into Embeddings

For each mention, we first obtain the mention representation m_i by passing it through the pre-trained RoBERTa model (Liu et al., 2019), which is a 12-layered transformer encoder. We transform this representation into an embedding using Equations 1 and 2 where W_w, b_w, W_r and b_r are trainable parameters. In order to avoid overfitting, we use dropout layer (Srivastava et al., 2014) to modify m_i and u_i.

$$u_i = tanh(W_w m_i + b_w) \tag{1}$$

$$r_i = W_r u_i + b_r \tag{2}$$

We train the model parameters using stochastic optimizer AdamW (Loshchilov and Hutter, 2018) which aims to minimize the cosine embedding loss shown in Equation 3 between the projected representation r_i and the corresponding target embedding t_i. We keep the target embeddings of the concepts fixed throughout the training.

$$loss = 1 - Cosine\ Similarity(r_i, t_i) \tag{3}$$

3. Datasets

CADEC: CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015) is one of the first publicly available datasets for MCN. It contains medical forum posts sourced from AskAPatient[3] which describe adverse drug events resulting from the usage of Diclofenac and Lipitor. These posts are largely written in colloquial language and often deviate from formal English grammar and punctuation rules. In this dataset, human annotators have marked the candidate medical concept mentions, such as *nt able 2 sleep 2nite*. Subsequently, each mention was manually mapped to a medical concept in the SNOMED-CT lexicon. Overall, it contains 6754 concept mentions which are mapped to 1029 SNOMED-CT concepts. Limsopatham and Collier (2016) randomly split this dataset into five folds and report the classification accuracy using five fold cross validation. However, Tutubalina et al. (2018); Miftahutdinov and Tutubalina (2019) discover that these folds contain nearly 66% redundant samples which causes high overlap between training and testing split, which in turn results in highly optimistic classification accuracy. In order to avoid this, Tutubalina et al. (2018) create custom folds without any overlap. These folds are made publicly available[4].

[3] https://www.askapatient.com

[4] https://yadi.sk/d/GZoWm1wBxzyW_w

PsyTAR: Psychiatric Treatment Adverse Reactions (PsyTAR) corpus (Zolnoori et al., 2019) is another publicly available dataset for MCN which contains 887 posts sourced from AskAPatient about psychiatric medications: Zoloft, Lexapro, Effexor and Cymbalta. In this dataset, 6556 concept mentions were extracted from these posts and were manually mapped to 618 SNOMED-CT codes. Miftahutdinov and Tutubalina (2019) discovered that this dataset also contains about 56% redundant examples and therefore random folds create a high overlap between train and test splits. In order to overcome this, they create custom folds with minimal overlap. These folds are also made publicly available[5].

SMM4H 2017: This dataset is part of a shared task, Social Media Mining for Health (SMM4H) which was organized by Sarker et al. (2018). It consists of mappings between user generated medical phrases extracted from Tweets and Medical Dictionary for Regulatory Activities (MedDRA) concepts. We do not use this dataset in this work as we restrict our attention to SNOMED-CT target concepts only.

SNOMED-CT Synonyms: The SNOMED-CT lexicon contains synonyms corresponding to each medical concept. We exploit this resource to generate labeled examples for MCN by treating each synonym as a medical concept mention. Table 1 depicts few example medical concepts and their synonyms.

Medical Concept ID	Concept Description	SNOMED-CT Synonyms
3424008	Tachycardia	rapid heart rate, increased heart rate
25064002	Pain in head	headache, head pain, cephalalgia, cephalgia, cephalodynia
68962001	Myalgia	muscle pain, muscle ache, myodynia, myosalgia, myoneuralgia
81680005	Cervicodynia	neck pain, painful neck, cervicalgia
18963009	Emotionally Labile	labile in mood, mood swing, variable mood, changeable mood, unstable mood, labile mood

Table 1: SNOMED-CT Medical Concepts and their Synonyms

4. Experimental Details

For the extent of this work, we use the custom folds released by Tutubalina et al. (2018) (for CADEC corpus) and Miftahutdinov and Tutubalina (2019) (for PsyTAR corpus). We

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

augment each train fold with labeled examples obtained from SNOMED-CT Synonyms dataset described in Section 3. We compute the classification accuracy for each of the five test folds and finally report the mean classification accuracy across all folds.

In order to obtain the optimal values for hyperparameters, we use 20% of training examples of each fold as validation set. We perform a random search over the hyperparameter space and choose the configuration which results in minimum validation loss. We also conduct experiments by keeping the weights of the pre-trained RoBERTa model fixed. We noticed that, this adversely affects the performance of our model.

5. Results

Method	**CADEC**	**PsyTAR**
GRU + Attention	66.56	65.98
GRU + Attention with Semantic features	70.05	68.59
ELMO + GRU + Attention	71.68	68.34
ELMO + GRU + Attention with semantic features	74.70	70.05
BERT	79.83	77.52
BERT with Semantic Features	79.25	77.33
RoBERTa with AvgEmb target embedding	79.34	79.76
RoBERTa with ELMo target embedding	80.94	80.36
RoBERTa with USE target embedding	81.62	81.16
RoBERTa with BERT target embedding	80.16	79.66
RoBERTa with Deepwalk target embedding	82.90	81.36
RoBERTa with Node2Vec target embedding	**83.18**	82.16
RoBERTa with LINE target embedding	82.44	**82.42**
RoBERTa with HARP target embedding	82.42	81.82

Table 2: The performance of the proposed approach and the state-of-the-art methods in terms of classification accuracy. The first six rows depict the performance of various approaches proposed by Miftahutdinov and Tutubalina (2019)

Table 2 details the comparative performance of our approach all the models proposed by the previous state-of-the-art (Miftahutdinov and Tutubalina, 2019) across CADEC and PsyTAR datasets. We observe almost all our models outperform the best models of Miftahutdinov and Tutubalina (2019). The highest performance of 83.18% and 82.42% was achieved by using graph base embeddings Node2Vec and LINE respectively. The previous state-of-the-art is a BERT based model which achieves a classification accuracy of 79.83%, 77.52% across CADEC and PsyTAR respectively. In comparison to this, our model demonstrates an improvement of 4.2%, 6.3% across these datasets. We observe that, the performance of graph embedding based methods: DeepWalk, Node2Vec, LINE and

HARP is superior to that of text embedding based methods. We also notice that Universal Sentence Encoder (USE) based model consistently performs better as compared to other text embedding based methods.

6. Analysis and Discussion

6.1. Failure Analysis

We discover that our model based on USE target embeddings is good at mapping mentions which have a lexical overlap with the target concept description. For instance, mapping *stomach discomfort* to `Stomach ache`. We observed that the model performance decreases when mapping mentions which do not contain any medical words. For instance, mapping *had me dodging parked cars like i 'm keanu reeves* to `hallucinations`. We also notice that our model wrongly maps mentions to the medical concepts whose embeddings are closer to the target embedding.

6.2. Semantic Similarity between Concept Mentions and Medical Concepts

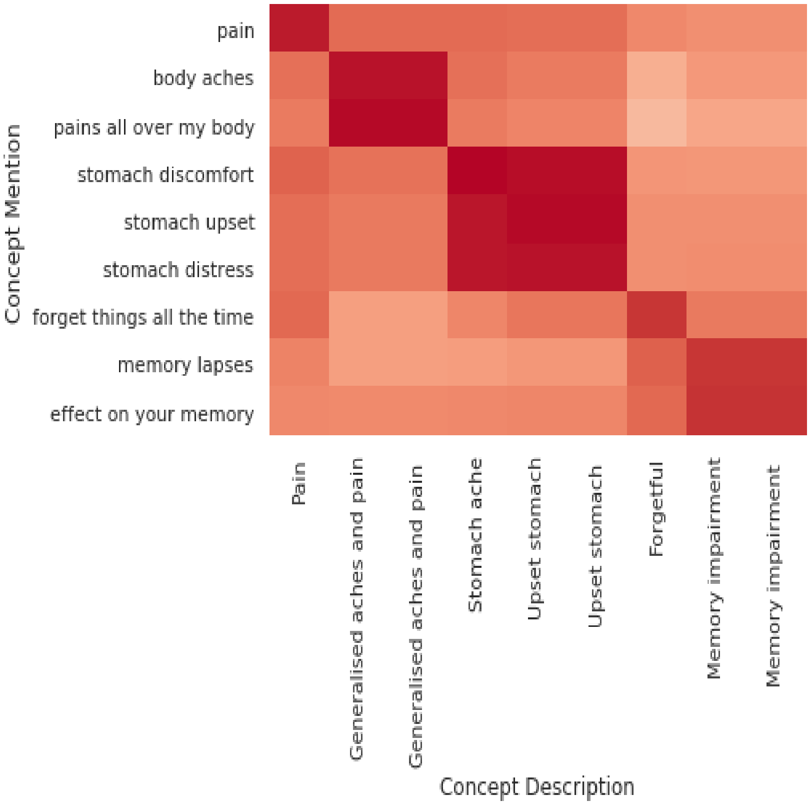

Figure 3: Heatmap depicting the strength of association between concept mentions and medical concept embeddings using USE based target embeddings.

Our model projects the concept mentions and the medical concepts to a common embedding space. As illustrated in Figure 3 these embeddings can be trivially used to visualize the semantic similarity between concept mentions and medical concepts.

6.3. Model performance when trained on SNOMED-CT Synonyms

Method	**CADEC**	**PsyTAR**
RoBERTa with LINE	59.2	56.0
RoBERTa with Node2Vec	62.9	52.8
RoBERTa with Deepwalk	63.2	53.4
RoBERTa with HARP	**64.8**	54.6
RoBERTa with USE	58.8	**58.4**
RoBERTa with BERT	47.5	39.9
RoBERTa with ELMO	53.8	46.0
RoBERTa with AvgEmb	58.3	53.0

Table 3: The performance of all proposed models across CADEC and PsyTAR datasets when trained on SNOMED-CT synonyms dataset.

In this study we wish to discover how well our model performs in mapping phrases to medical concepts when the human labeled mappings (such as *not able to sleep 2nite* and `Insomnia`) are not present in the training set. We therefore conduct experiments, where we train our model using SNOMED-CT Synonyms only and evaluate the model on CADEC and PsyTAR datasets. Table 3 demonstrates the performance of our approach across these datasets. We find that our model based on HARP embeddings and USE embeddings achieve a classification accuracy of 64.8% and 58.4% on these datasets. We also observe that the classification accuracy in case of CADEC dataset is comparable to the GRU + Attention based model proposed by the current state-of-the-art (Miftahutdinov and Tutubalina, 2019) (refer Table 2).

6.4. Model performance on Medical Concepts Not Encountered during Training

In this study we wish to discover how well our model performs in mapping phrases to medical concepts which were not present in the training set. We create two groups of mutually exclusive medical concepts. We train our model on the samples whose target concepts are present in the first group and evaluate it on examples whose target concept belong to the second group. As expected, we observed that this adversely affects the performance of our models. However, we noticed that our models are able to fetch the correct target concept amongst the top ten predictions. Table 4 shows the performance (Recall at rank 10) of the proposed models. We find that Deepwalk and USE based embeddings outperform other models on CADEC and PsyTAR datasets respectively. We also notice that the LINE based embedding method consistently underperforms compared to other methods. We wish to

Method	CADEC	PsyTAR
RoBERTa with LINE	13.23	02.27
RoBERTa with Node2Vec	73.52	28.88
RoBERTa with Deepwalk	**75.00**	23.40
RoBERTa with HARP	73.52	27.65
RoBERTa with USE	39.70	**57.44**
RoBERTa with BERT	20.58	29.78
RoBERTa with ELMO	55.88	55.32
RoBERTa with AvgEmb	26.47	42.55

Table 4: The Recall at Rank 10 of all proposed models across CADEC and PsyTAR datasets. The target labels were not encountered during training phase.

investigate the cause for this phenomenon in future. Our experiments show that although we cannot completely eliminate the need for labeled examples, our formulation helps minimize the labeling effort.

7. Conclusions and Future Work

In this work, we propose a novel method for Medical Concept Normalization which maps a medical concept mention to its corresponding medical concept in a target lexicon. Unlike previous works, our method scales to millions of concepts and even handles increase in target lexicon size. Experiments on two major datasets show that our model outperforms the previous state-of-the-art MCN models. We conduct several studies to study the robustness of our model under different training conditions.

In this work, we explored two types of methods to encode target knowledge - text and graph embedding based methods. We reported the performance of eight different target encoding methods. We see a great scope for improvement using ensemble methods where multiple embedding types can be combined to obtain heterogeneous concept embeddings. Our current implementation does not take into account the nature of the relationship between concepts. In future, we would like to overcome this problem by encoding the target knowledge using Knowledge Base embeddings. Finally, we would also like to experiment with other deep learning based approaches to encode SNOMED-CT graph.

References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 819–826, 2013.

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

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Deanna J Attai, Michael S Cowher, Mohammed Al-Hamadani, Jody M Schoger, Alicia C Staley, and Jeffrey Landercasper. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. *Journal of medical Internet research*, 17(7):e188, 2015.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

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. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. Harp: Hierarchical representation learning for networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Christine C Cheston, Tabor E Flickinger, and Margaret S Chisolm. Social media use in medical education: a systematic review. *Academic Medicine*, 88(6):893–901, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78–94, 2018.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28(2):361–368, 2009.

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

Nut Limsopatham and Nigel Collier. Adapting phrase-based machine translation to normalise medical terms in social media messages. *arXiv preprint arXiv:1508.02285*, 2015.

Nut Limsopatham and Nigel Collier. 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)*, volume 1, pages 1014–1023, 2016.

Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Multi-label classification via feature-aware implicit label space encoding. In *International conference on machine learning*, pages 325–333, 2014.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. 2018.

Zulfat Miftahutdinov and Elena Tutubalina. Deep neural models for medical concept normalization in user-generated texts. *arXiv preprint arXiv:1907.07972*, 2019.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44, 2013.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters*, 49(3):1239–1256, 2019.

Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific, 2016.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, et al. 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, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077. International World Wide Web Conferences Steering Committee, 2015.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102, 2018.

C Lee Ventola. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics*, 39(7):491, 2014.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*, 2018.

Henghui Zhu, Ioannis Ch Paschalidis, and Amir Tahmasebi. Clinical concept extraction with contextual word embedding. *arXiv preprint arXiv:1810.10566*, 2018.

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. 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, 2019.