Deep Neural Models for Medical Concept Normalization in User-Generated Texts

Zulfat Miftahutdinov
Kazan Federal University,
Kazan, Russia
zulfatmi@gmail.com

Elena Tutubalina
Kazan Federal University,
Kazan, Russia
Samsung-PDMI Joint AI Center,
PDMI RAS, St. Petersburg, Russia
elvtutubalina@kpfu.ru

Abstract

In this work, we consider the *medical concept normalization* problem, i.e., the problem of mapping a health-related entity mention in a free-form text to a concept in a controlled vocabulary, usually to the standard thesaurus in the Unified Medical Language System (UMLS). This is a challenging task since medical terminology is very different when coming from health care professionals or from the general public in the form of social media texts. We approach it as a sequence learning problem with powerful neural networks such as recurrent neural networks and contextualized word representation models trained to obtain semantic representations of social media expressions. Our experimental evaluation over three different benchmarks shows that neural architectures leverage the semantic meaning of the entity mention and significantly outperform an existing state of the art models.

1 Introduction

User-generated texts (UGT) on social media present a wide variety of facts, experiences, and opinions on numerous topics, and this treasure trove of information is currently severely underexplored. We consider the problem of discovering medical concepts in UGTs with the ultimate goal of mining new symptoms, adverse drug reactions (ADR), and other information about a disorder or a drug.

An important part of this problem is to translate a text from "social media language" (e.g., "can't fall asleep all night" or "head spinning a little") to "formal medical language" (e.g., "insomnia" and "dizziness" respectively). This is necessary to match user-generated descriptions with medical concepts, but it is more than just a simple matching of UGTs against a vocabulary. We call the task of mapping the language of UGTs to medical terminology *medical concept normalization*. It is especially difficult since in social media, patients discuss different concepts of illness and a wide array of drug reactions. Moreover, UGTs from social networks are typically ambiguous and very noisy, containing misspelled words, incorrect grammar, hashtags, abbreviations, smileys, different variations of the same word, and so on.

Traditional approaches for concept normalization utilized lexicons and knowledge bases with string matching. The most popular knowledge-based system for mapping texts to UMLS identifiers is MetaMap (Aronson, 2001). This linguistic-based system uses lexical lookup and variants by associating a score with phrases in a sentence. The state-of-the-art baseline for clinical and scientific texts is DNorm (Leaman et al., 2013). DNorm adopts a pairwise learning-to-rank technique using vectors of query mentions and candidate concept terms. This model outperforms MetaMap significantly, increasing the macro-averaged F-measure by 25% on an NCBI disease dataset. However, while these tools have proven to be effective for patient records and research papers, they achieve moderate results on social media texts (Nikfarjam et al., 2015; Limsopatham and Collier, 2016).

Recent works go beyond string matching: these works have tried to view the problem of matching a one- or multi-word expression against a knowledge base as a supervised sequence labeling problem. Limsopatham and Collier (2016) utilized convolutional neural networks (CNNs) for phrase normalization in user reviews, while Tutubalina et al. (2018), Han et al. (2017), and Belousov et al. (2017) applied recurrent neural networks (RNNs) to UGTs, achieving similar results. These works were among the first applications of deep learning techniques to medical concept normalization.

The goal of this work is to study the use of deep neural models, i.e., contextualized word represen-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 393–399
Florence, Italy, July 28 - August 2, 2019.

Entity from UGTs	Medical Concept
no sexual interest	Lack of libido
nonsexual being	Lack of libido
couldnt remember long periods of time or things	Poor long-term memory
loss of memory	Amnesia
bit of lower back pain	Low Back Pain
pains	Pain
like i went downhill	Depressed mood
just lived day by day	Apathy
dry mouth	Xerostomia

Table 1: Examples of extracted social media entities and their associated medical concepts.

tation model BERT (Devlin et al., 2018) and Gated Recurrent Units (GRU) (Cho et al., 2014) with an attention mechanism, paired with *word2vec* word embeddings and contextualized ELMo embeddings (Peters et al., 2018). We investigate if a joint architecture with special provisions for domain knowledge can further improve the mapping of entity mentions from UGTs to medical concepts. We combine the representation of an entity mention constructed by a neural model and distance-like similarity features using vectors of an entity mention and concepts from the UMLS. We experimentally demonstrate the effectiveness of the neural models for medical concept normalization on three real-life datasets of tweets and user reviews about medications with two evaluation procedures.

2 Problem Statement

Our main research problem is to investigate the content of UGTs with the aim to learn the transition between a laypersons language and formal medical language. Examples from Table 1 show that an automated model has to account for the semantics of an entity mention. For example, it has to be able to map not only phases with shared n-grams *no sexual interest* and *nonsexual being* into the concept "Lack of libido" but also separate the phase *bit of lower back pain* from the broader concept "Pain" and map it to a narrower concept.

While focusing on user-generated texts on social media, in this work we seek to answer the following research questions.

RQ1: Do distributed representations reveal important features for medication use in user-generated texts?

RQ2: Can we exploit the semantic similarity between entity mentions from user comments and medical concepts? Do the neural models produce better results than the existing effective baselines? [current research]

RQ3: How to integrate linguistic knowledge about concepts into the models? [current research]

RQ4: How to jointly learn concept embeddings from UMLS and representations of health-related entities from UGTs? [future research]

RQ5: How to effectively use of contextual information to map entity mentions to medical concepts? [future research]

To answer RQ1, we began by collecting UGTs from popular medical web portals and investigating distributed word representations trained on 2.6 millions of health-related user comments. In particular, we analyze drug name representations using clustering and chemoinformatics approaches. The analysis demonstrated that similar word vectors correspond to either drugs with the same active compound or to drugs with close therapeutic effects that belong to the same therapeutic group. It is worth noting that chemical similarity in such drug pairs was found to be low. Hence, these representations can help in the search for compounds with potentially similar biological effects among drugs of different therapeutic groups (Tutubalina et al., 2017).

To answer RQ2 and RQ3, we develop several models and conduct a set of experiments on three benchmark datasets where social media texts are extracted from user reviews and Twitter. We present this work in Sections 3 and 4. We discuss RQ4 and RQ5 with research plans in Section 5.

3 Methods

Following state-of-the-art research (Limsopatham and Collier, 2016; Sarker et al., 2018), we view concept normalization as a classification problem.

To answer RQ2, we investigate the use of neural networks to learn the semantic representation of an entity before mapping its representation to a medical concept. First, we convert each mention into a vector representation using one of the following (well-known) neural models:

(1) bidirectional LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014) with an attention mechanism and a hyperbolic tangent activation function on top of 200-dimensional word embeddings obtained to answer RQ1;

(2) a bidirectional layer with attention on top of deep contextualized word representations ELMo (Peters et al., 2018);

(3) a contextualized word representation model BERT (Devlin et al., 2018), which is a multi-layer bidirectional Transformer encoder.

We omit technical explanations of the neural network architectures due to space constraints and refer to the studies above.

Next, the learned representation is concatenated with a number of semantic similarity features based on prior knowledge from the UMLS Metathesaurus. Lastly, we add a softmax layer to convert values to conditional probabilities.

The most attractive feature of the biomedical domain is that domain knowledge is prevailing in this domain for dozens of languages. In particular, UMLS is undoubtedly the largest lexico-semantic resource for medicine, containing more than 150 lexicons with terms from 25 languages. To answer RQ3, we extract a set of features to enhance the representation of phrases. These features contain cosine similarities between the vectors of an input phrase and a concept in a medical terminology dictionary. We use the following strategy, which we call TF-IDF (MAX), to construct representations of a concept and a mention: represent a medical code as a set of terms; for each term, compute the cosine distance between its TF-IDF representation and the entity mention; then choose the term with the largest similarity.

4 Experiments

We perform an extensive evaluation of neural models on three datasets of UGTs, namely **CADEC** (Karimi et al., 2015), **PsyTAR** (Zolnoori et al., 2019), and **SMM4H 2017** (Sarker et al., 2018). The basic task is to map a social media phrase to a relevant medical concept.

4.1 Data

CADEC. CSIRO Adverse Drug Event Corpus (CADEC) (Karimi et al., 2015) is the first richly annotated and publicly available corpus of medical forum posts taken from *AskaPatient*[1]. This dataset contains 1253 UGTs about 12 drugs divided into two categories: Diclofenac and Lipitor. All posts were annotated manually for 5 types of entities: ADR, Drug, Disease, Symptom, and Finding. The annotators performed terminology association using the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED CT). We removed "conceptless" or ambiguous mentions for the purposes of evaluation. There were 6,754 entities and 1,029 unique codes in total.

PsyTAR. Psychiatric Treatment Adverse Reactions (PsyTAR) corpus (Zolnoori et al., 2019) is the second open-source corpus of user-generated posts taken from AskaPatient. This dataset includes 887 posts about four psychiatric medications from two classes: (i) Zoloft and Lexapro from the Selective Serotonin Reuptake Inhibitor (SSRI) class and (ii) Effexor and Cymbalta from the Serotonin Norepinephrine Reuptake Inhibitor (SNRI) class. All posts were annotated manually for 4 types of entities: ADR, withdrawal symptoms, drug indications, and sign/symptoms/illness. The corpus consists of 6556 phrases mapped to 618 SNOMED codes.

SMM4H 2017. In 2017, Sarker et al. (2018) organized the Social Media Mining for Health (SMM4H) shared task which introduced a dataset with annotated ADR expressions from *Twitter*. Tweets were collected using 250 keywords such as generic and trade names for medications along with misspellings. Manually extracted ADR expressions were mapped to Preferred Terms (PTs) of the Medical Dictionary for Regulatory Activities (MedDRA). The training set consists of 6650 phrases mapped to 472 PTs. The test set consists of 2500 mentions mapped to 254 PTs.

4.2 Evaluation Details

We evaluate our models based on classification accuracy, averaged across randomly divided five folds of the CADEC and PsyTAR corpora. For SMM4H 2017 data, we adopted the official training and test sets (Sarker et al., 2018). Analysis of randomly split folds shows that *Random KFolds* create a high overlap of expressions in exact matching between subsets (see the baseline results in Table 2). Therefore, we set up a

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

specific train/test split procedure for 5-fold cross-validation on the CADEC and PsyTAR corpora: we removed duplicates of mentions and grouped medical records we are working with into sets related to specific medical codes. Then, each set has been split independently into k folds, and all folds have been merged into the final k folds named *Custom KFolds*. Random folds of CADEC are adopted from (Limsopatham and Collier, 2016) for a fair comparison. Custom folds of CADEC are adopted from our previous work (Tutubalina et al., 2018). PsyTAR folds are available on Zenodo.org[2]. We have also implemented a simple *baseline* approach that uses exact lexical matching with lowercased annotations from the training set.

4.3 Results

Table 2 shows our results for the concept normalization task on the Random and Custom KFolds of the CADEC, PsyTAR, and SMM4H 2017 corpora.

To answer RQ2, we compare the performance of examined neural models with the baseline and state-of-the-art methods in terms of accuracy. Attention-based GRU with ELMo embeddings showed improvement over GRU with *word2vec* embeddings, increasing the average accuracy to 77.85 (+3.65). The semantic information of an entity mention learned by BERT helps to improve the mapping abilities, outperforming other models (avg. accuracy 83.67). Our experiments with recurrent units showed that GRU consistently outperformed LSTM on all subsets, and attention mechanism provided further quality improvements for GRU. From the difference in accuracy on the Random and Custom KFolds, we conclude that future research should focus on developing extrinsic test sets for medical concept normalization. In particular, the BERT model's accuracy on the CADEC Custom KFolds decreased by 9.23% compared to the CADEC Random KFolds.

To answer RQ3, we compare the performance of models with additional similarity features (marked by "w/") with others. Indeed, joint models based on GRU and similarity features gain 2-5% improvement on sets with Custom KFolds. The joint model based on BERT and similarity features stays roughly on par with BERT on all sets. We also tested different strategies for constructing representations using word embeddings and TF-IDF for all synonyms' tokens that led to similar improvements for GRU.

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

5 Future Directions

RQ4. Future research might focus on developing an embedding method that jointly maps extracted entity mentions and UMLS concepts into the same continuous vector space. The methods could help us to easily measure the similarity between words and concepts in the same space. Recently, Yamada et al. (2016) demonstrated that co-trained vectors improve the quality of both word and entity representations in entity linking (EL) which is a task closely related to concept normalization. We note that most of the recent EL methods focus on the disambiguation sub-task, applying simple heuristics for candidate generation. The latter is especially challenging in medical concept normalization due to a significant language difference between medical terminology and patient vocabulary.

RQ5. Error analysis has confirmed that models often misclassify closely related concepts (e.g., "Emotionally detached" and "Apathy") and antonymous concepts (e.g., "Hypertension" and "Hypotension"). We suggest to take into account not only the distance-like similarity between entity mentions and concepts but the mention's context, which is not used directly in recent studies on concept normalization. The context can be represented by the set of adjacent words or entities. As an alternative, one can use a conditional random field (CRF) to output the most likely sequence of medical concepts discussed in a review.

6 Related Work

In 2004, the research community started to address the needs to automatically detect biomedical entities in free texts through shared tasks. Huang and Lu (2015) survey the work done in the organization of biomedical NLP (BioNLP) challenge evaluations up to 2014. These tasks are devoted to the normalization of (1) genes from scientific articles (BioCreative I-III in 2005-2011); (2) chemical entity mentions (BioCreative IV CHEMDNER in 2014); (3) disorders from abstracts (BioCreative V CDR Task in 2015); (4) diseases from clinical reports (ShARe/CLEF eHealth 2013; SemEval 2014 task 7). Similarly, the *CLEF Health* 2016

Method	CADEC		PsyTAR		SMM4H
	Random	Custom	Random	Custom	Official
Baseline: match with training set annotation	66.09	0.0	56.04	2.63	67.12
DNorm (Limsopatham and Collier, 2016)	73.39	-	-	-	-
CNN (Limsopatham and Collier, 2016)	81.41	-	-	-	-
RNN (Limsopatham and Collier, 2016)	79.98	-	-	-	-
Attentional Char-CNN (Niu et al., 2018)	84.65	-	-	-	-
Hierarchical Char-CNN (Han et al., 2017)	-	-	-	-	87.7
Ensemble (Sarker et al., 2018)	-	-	-	-	88.7
GRU+Attention	82.19	66.56	73.12	65.98	83.16
GRU+Attention w/ TF-IDF (MAX)	84.23	70.05	75.53	68.59	86.28
ELMo+GRU+Attention	85.06	71.68	77.58	68.34	86.60
ELMo+GRU+Attention w/ TF-IDF (MAX)	85.71	74.70	79.52	70.05	87.52
BERT	88.69	79.83	83.07	77.52	89.28
BERT w/ TF-IDF (MAX)	88.84	79.25	82.37	77.33	89.64

Table 2: The performance of the proposed models and the state-of-the-art methods in terms of accuracy.

and 2017 labs addressed the problem of ICD coding of free-form death certificates (without specified entity mentions). Traditionally, linguistic approaches based on dictionaries, association measures, and syntactic properties have been used to map texts to a concept from a controlled vocabulary (Aronson, 2001; Van Mulligen et al., 2016; Mottin et al., 2016; Ghiasvand and Kate, 2014; Tang et al., 2014). Leaman et al. (2013) proposed the DNorm system based on a pairwise learning-to-rank technique using vectors of query mentions and candidate concept terms. These vectors are obtained from a tf-idf representation of all tokens from training mentions and concept terms. Zweigenbaum and Lavergne (2016) utilized a hybrid method combining simple dictionary projection and mono-label supervised classification from ICD coding. Nevertheless, the majority of biomedical research on medical concept extraction primarily focused on scientific literature and clinical records (Huang and Lu, 2015). Zolnoori et al. (2019) applied a popular dictionary look-up system cTAKES on user reviews. cTAKES based on additional PsyTAR's dictionaries achieves twice better results (0.49 F1 score on the exact matching). Thus, dictionaries gathered from layperson language can efficiently improve automatic performance.

The 2017 SMM4H shared task (Sarker et al., 2018) was the first effort for the evaluation of NLP methods for the normalization of health-related text from social media on publicly released data. Recent advances in neural networks have been utilized for concept normalization: recent studies have employed convolutional neural networks (Limsopatham and Collier, 2016; Niu et al., 2018) and recurrent neural networks (Belousov et al., 2017; Han et al., 2017). These works have trained neural networks from scratch using only entity mentions from training data and pre-trained word embeddings. To sum up, most methods have dealt with encoding information an entity mention itself, ignoring the broader context where it occurred. Moreover, these studies did not examine an evaluation methodology tailored to the task.

7 Conclusion

In this work, we have performed a fine-grained evaluation of neural models for medical concept normalization tasks. We employed several powerful models such as BERT and RNNs paired with pre-trained word embeddings and ELMo embeddings. We also developed a joint model that combines (i) semantic similarity features based on prior knowledge from UMLS and (ii) a learned representation that captures extensional semantic information of an entity mention. We have carried out experiments on three datasets using 5-fold cross-validation in two setups. Each dataset contains phrases and their corresponding SNOMED or MedDRA concepts. Analyzing the results, we have found that similarity features help to improve mapping abilities of joint models based on recurrent neural networks paired with pre-trained word embeddings or ELMo embeddings while staying roughly on par with the advanced language repre-

sentation model BERT in terms of accuracy. Different setups of evaluation procedures affect the performance of models significantly: the accuracy of BERT is 7.25% higher on test sets with a simple random split than on test sets with the proposed custom split. Moreover, we have discussed some interesting future research directions and challenges to be overcome.

Acknowledgments

We thank Sergey Nikolenko for helpful discussions. This research was supported by the Russian Science Foundation grant no. 18-11-00284.

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.

M. Belousov, W. Dixon, and G. Nenadic. 2017. Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task. *CEUR Workshop Proceedings*, 1996:54–58.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

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

Omid Ghiasvand and Rohit J Kate. 2014. Uwm: Disorder mention extraction from clinical text using crfs and normalization using learned edit distance patterns. In *SemEval@ COLING*, pages 828–832.

S. Han, T. Tran, A. Rios, and R. Kavuluru. 2017. Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter. *CEUR Workshop Proceedings*, 1996:49–53.

S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. Based on TR FKI-207-95, TUM (1995).

Chung-Chi Huang and Zhiyong Lu. 2015. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1):132–144.

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.

Nut Limsopatham and Nigel Collier. 2016. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *ACL*.

Luc Mottin, Julien Gobeill, Anaïs Mottaz, Emilie Pasche, Arnaud Gaudinat, and Patrick Ruch. 2016. Bitem at clef ehealth evaluation lab 2016 task 2: Multilingual information extraction. In *CLEF (Working Notes)*, pages 94–102.

Azadeh Nikfarjam, Abeed Sarker, Karen OConnor, Rachel Ginn, and Graciela Gonzalez. 2015. 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.

Jinghao Niu, Yehui Yang, Siheng Zhang, Zhengya Sun, and Wensheng Zhang. 2018. Multi-task character-level attentional networks for medical concept normalization. *Neural Processing Letters*, pages 1–18.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

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.

Yaoyun Zhang1 Jingqi Wang1 Buzhou Tang, Yonghui Wu1 Min Jiang, and Yukun Chen3 Hua Xu. 2014. Uth_ccb: a report for semeval 2014–task 7 analysis of clinical text. *SemEval 2014*, page 802.

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.

EV Tutubalina, Z Sh Miftahutdinov, RI Nugmanov, TI Madzhidov, SI Nikolenko, IS Alimova, and AE Tropsha. 2017. Using semantic analysis of texts for the identification of drugs with similar therapeutic effects. *Russian Chemical Bulletin*, 66(11):2180–2189.

E Van Mulligen, Zubair Afzal, Saber A Akhondi, Dang Vo, and Jan A Kors. 2016. Erasmus MC at CLEF eHealth 2016: Concept recognition and coding in French texts. CLEF.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259.

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.

Pierre Zweigenbaum and Thomas Lavergne. 2016. Hybrid methods for icd-10 coding of death certificates. *EMNLP 2016*, page 96.