Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation

Nut Limsopatham and **Nigel Collier**
Language Technology Lab
Department of Theoretical and Applied Linguistics
University of Cambridge
Cambridge, UK
{nl347,nhc30}@cam.ac.uk

Abstract

Automatically recognising medical concepts mentioned in social media messages (e.g. tweets) enables several applications for enhancing health quality of people in a community, e.g. real-time monitoring of infectious diseases in population. However, the discrepancy between the type of language used in social media and medical ontologies poses a major challenge. Existing studies deal with this challenge by employing techniques, such as lexical term matching and statistical machine translation. In this work, we handle the medical concept normalisation at the semantic level. We investigate the use of neural networks to learn the transition between layman's language used in social media messages and formal medical language used in the descriptions of medical concepts in a standard ontology. We evaluate our approaches using three different datasets, where social media texts are extracted from Twitter messages and blog posts. Our experimental results show that our proposed approaches significantly and consistently outperform existing effective baselines, which achieved state-of-the-art performance on several medical concept normalisation tasks, by up to 44%.

1 Introduction

Existing studies (O'Connor et al., 2014; Limsopatham and Collier, 2015a; Limsopatham and Collier, 2015b) have shown that data from social media (e.g. Twitter[1] and Facebook[2]) can be leveraged to improve the understanding of patients' experience in healthcare, such as the spread of infectious diseases and side-effects of drugs. However, the lexical and grammatical variability of the language used in social media poses a key challenge for extracting information (Baldwin et al., 2013; O'Connor et al., 2014). In particular, the frequent use of informal language, non-standard grammar and abbreviation forms, as well as typos in social media messages has to be taken into account by effective information extraction systems.

The task of medical concept normalisation for social media text, which aims to map a variable length social media message to a medical concept in some external coding system, is faced with a similar challenge (Limsopatham and Collier, 2015b). Traditional approaches, e.g. (Ristad and Yianilos, 1998; Aronson, 2001; Lu et al., 2011; McCallum et al., 2012), used proximity matching or heuristic string matching rules based on dictionary lookup when mapping texts to medical concepts. For example, Ristad and Yianilos (1998) incorporated edit-distance when mapping similar texts. The MetaMap system of Aronson (2001) applied a rule-based approach using pre-defined variants of terms when mapping texts to medical concepts in the UMLS Metathesaurus[3]. However, as shown in Table 1, existing string matching techniques may not be able to map the social media message "moon face and 30 lbs in 6 weeks" to the medical concept 'Weight Gain', or map "head spinning a little" to 'Dizziness', as no words in the social media messages and the description of the medical concepts correspond. Recent studies, e.g. (Leaman et al., 2013; Leaman and Lu, 2014; Limsopatham and Collier, 2015a), applied machine learning techniques to take into account relationships between different words (e.g. synonyms) when performing normal-

[1] http://twitter.com
[2] http://facebook.com
[3] https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1014–1023,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

Social media message	Description of corresponding medical concept
lose my appetite	Loss of appetite
i don’t hunger or thirst	Loss of Appetite
hungry	Hunger
moon face and 30 lbs in 6 weeks	Weight Gain
gained 7 lbs	Weight Gain
lose the 10 lbs	Body Weight Decreased
feeling dizzy ...	Dizziness
head spinning a little	Dizziness
terrible headache!!	Headache

Table 1: Examples of social media messages and their related medical concepts.

isation. For instance, the *DNorm* system of Leaman et al. (2013), which achieved state-of-the-art performance on several medical concept normalisation tasks for medical articles (Doğan et al., 2014) and patient records (Suominen et al., 2013), used a pairwise learning-to-rank technique to learn the similarity between different terms when performing concept normalisation. Limsopatham and Collier (2015a) leveraged translations between the informal language used in social media and the formal language used in the description of medical concepts in an ontology. However, we argue that effective concept normalisation requires a system to take into account the semantics of social media messages and medical concepts. For example, to be able to map from the social media message “i don’t hunger or thirst” to the medical concept ‘Loss of Appetite’, a normalisation system has to take into account the semantics of the whole message; otherwise, “i don’t hunger or thirst” may be mapped to the medical concept ‘Hunger’, because they contain the term “hunger” in common.

In this work, we go beyond string matching. We propose to learn and exploit the semantic similarity between texts from social media messages and medical concepts using deep neural networks. In particular, we investigate the use of techniques from two families of deep neural networks, i.e. a convolutional neural network (CNN) and a recurrent neural network (RNN), to learn the mapping between social media texts and medical concepts. We evaluate our approaches using three different datasets that contain messages from Twitter and blog posts. Our experimental results show that our proposed approaches significantly outperform existing strong baselines (e.g. DNorm) across all of the three datasets. The performance improvement is by up to 44%.

The main contributions of this paper are three-fold:

1. We propose two novel approaches based on CNN and RNN for medical concept normalisation.
2. We introduce two datasets with the gold-standard mappings between medical concepts and social media texts extracted from tweets and blog posts, respectively.
3. We thoroughly evaluate our proposed approaches using these two datasets and an existing dataset of tweets related to the topic of adverse drug reactions (ADRs) (Limsopatham and Collier, 2015a).

The remainder of this paper is organised as follows. In Section 2, we discuss related work and position our paper in the literature. Section 3 introduces our neural network approaches for medical concept normalisation. We describe our experimental setup and empirically evaluate our proposed approaches in Sections 4 and 5, respectively. We provide further analysis and discussion of our approaches in Section 6. Finally, Section 7 provides concluding remarks.

2 Related Work

Existing techniques for concept normalisation are mostly based on string matching (e.g. (Tsuruoka et al., 2007; Ristad and Yianilos, 1998; Lu et al., 2011; McCallum et al., 2012). For example, McCallum et al. (2012) used conditional random field to learn edit distance between phrases. In the medical domain, Tsuruoka et al. (2007) learned mappings between phrases in medical documents and medical concepts by using string matching features, such as character bigrams and common tokens. Meanwhile, Metke-Jimenez and Karimi (2015), and O'Connor et al. (2014) used term weighting techniques, such as TF-IDF and BM25 (Robertson and Zaragoza, 2009) to retrieve relevant concepts. We tackle the concept normalisation task in a different manner. In particular, we use deep neural networks to capture the similarity and/or dependency between terms and effectively represent a given social media message in a low dimensional vector representation, before mapping it to a medical concept.

Another research area related to this work is the exploitation of word embeddings (i.e. distributed vector representation of words). It has been empirically shown that word embeddings can capture semantic and syntactic similarities between words (Turian et al., 2010; Mikolov et al., 2013b; Pennington et al., 2014; Levy and Goldberg, 2014). The cosine similarity between vectors of words has a positive correlation with the semantic similarity between them (Mikolov et al., 2013b; Pennington et al., 2014). Importantly, word embeddings have been effectively used for several NLP tasks, such as named entity recognition (Passos et al., 2014), machine translation (Mikolov et al., 2013a) and part-of-speech tagging (Turian et al., 2010). In the context of concept normalisation, Limsopatham and Collier (2015a) showed that effective performance could be achieved by mapping the processed social media messages and medical concepts using the similarity of their embeddings. In this work, we use word embeddings as inputs of deep neural networks, which would allow an effective representation of words when learning the concept normalisation.

Neural networks, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), have been effectively applied to NLP tasks, such as NER, sentiment classifications and machine translation (Collobert et al., 2011; Kim, 2014; Bahdanau et al., 2014). For example, Collobert et al. (2011) effectively used a multilayer neural network for chunking, part-of-speech tagging, NER and semantic role labelling. Kim (2014) effectively used CNN with pre-built word embeddings when performing sentence classifications. Kalchbrenner et al. (2014) learned representation of sentences by using CNN. Meanwhile, Bahdanau et al. (2014) used RNN to encode a sentence written in one language (e.g. French) into a fixed length vector before decoding it to a sentence in another language (e.g. English) for translation. Socher et al. used recursive neural networks to model sentences for different tasks, including paraphrase detection (Socher et al., 2011) and sentence classification (Socher et al., 2013). In this paper, we investigate only the use of CNN and RNN for medical concept normalisation, as recursive neural networks require parse trees of input sentences while grammatical rules are typically ignored in social media messages.

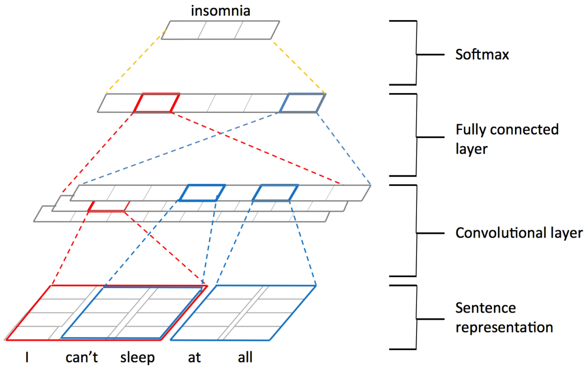


Figure 1: Our CNN architecture for medical concept normalisation.

3 Neural Networks for Concept Normalisation

Next, we introduce our medical concept normalisation approaches based on CNN and RNN in Sections 3.1 and 3.2, respectively.

3.1 CNN for Concept Normalisation

Our first approach uses CNN to learn the semantic representation of a social media message before mapping it to an appropriate medical concept. We use a CNN architecture with a single convolutional and pooling layer, as shown in Figure 1. Specifically, we firstly represent a given social media message of length l words (padded where necessary) using a sentence matrix $\mathbf{S} \in \mathbb{R}^{d \times l}$:

$$\mathbf{S} = \begin{bmatrix} | & | & | & | & \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \dots & \mathbf{x}_l \\ | & | & | & | & \end{bmatrix} \quad (1)$$

where each column of $\mathbf{S}$ is the d-dimensional vector (i.e. embedding) $\mathbf{x}_i \in \mathbb{R}^d$ of each word in the social media message, which can be retrieved from pre-built word embeddings. This allows the model to take into account semantic features from the embeddings of each word.

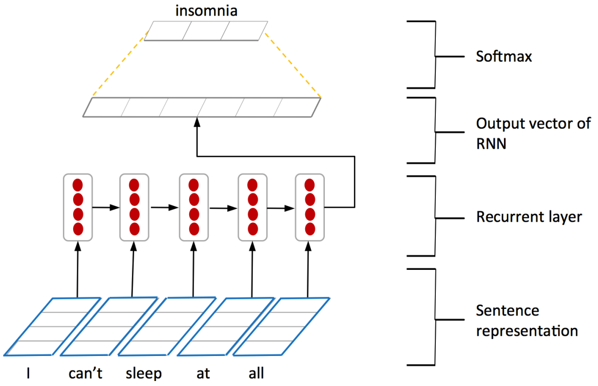


Figure 2: Our RNN architecture for medical concept normalisation.

We then apply a convolution operation using a filter $\mathbf{w} \in \mathbb{R}^{d \times h}$ to a window of h words. In particular, the filter $\mathbf{w}$ is convolved over the sequence of words in the sentence matrix $\mathbf{S}$ to create a feature matrix $\mathbf{C}$. Each feature c_i in $\mathbf{C}$ is extracted from a window of words $\mathbf{x}_{i:i+h-1}$, as follow:

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b) \tag{2}$$

where f is an activation function, such as sigmoid or tanh, and $b \in \mathbb{R}$ is a bias. Note that multiple filters (e.g. using different size h of window of words) can be used to extract multiple features.

This convolution operation enables the learning of dependencies between words from their semantic representation (i.e. word embeddings). In order to capture the most important features, max pooling (Collobert et al., 2011) is applied to take the maximum value of each row in the matrix $\mathbf{C}$:

$$\mathbf{c}_{max} = \begin{bmatrix} max(\mathbf{C}_{1,:}) \\ \vdots \\ max(\mathbf{C}_{d,:}) \end{bmatrix} \tag{3}$$

Finally, the fixed sized vector $\mathbf{c}_{max}$ forms a fully connected layer, which is used as inputs of softmax for multi-class classification. Indeed, the vector $\mathbf{c}_{max}$ provides a sentence representation that captures an extensional semantic information of the social media message for softmax to map to an appropriate medical concept.

3.2 RNN for Concept Normalisation

Our second approach uses RNN to capture the semantics of sequences of words in a social media message during normalisation. This approach is different from the CNN approach (introduced Section 3.1) in that instead of using the convolutional layer to learn the representation of social media messages (i.e. the vector representation at the fully connected layer), our RNN approach deploys a recurrent layer, as shown in Figure 2. Similar to the CNN approach, we initially represent a social media message of length l words using a sentence matrix $\mathbf{S} \in \mathbb{R}^{d \times l}$, as in Equation (1).

	TwADR-S	TwADR-L	AskAPatient
$\lvert Q \rvert$	201	1,436	8,662
$\lvert V_Q \rvert$	488	995	2,872
$\lvert C \rvert$	58	2,200	1,036
$\lvert V_C \rvert$	98	2,394	1,200
$\lvert Q \mapsto C \rvert_{avg}$	3.4655	0.6428	8.3610
$\lvert Q \mapsto C \rvert_{SD}$	5.6264	3.3168	39.2009
$\lvert Q \mapsto C \rvert_{min}$	1	0	1
$\lvert Q \mapsto C \rvert_{max}$	35	58	1,073

Table 2: Statistics of the datasets used in the experiments. $|Q|$: Number of queries. $|V_Q|$: Vocabulary size of queries. $|C|$: Number of target concepts. $|V_C|$: Vocabulary size of definition of target concepts. $|Q \mapsto C|_{avg}$ and $|Q \mapsto C|_{SD}$: Average number of queries mapped to each target concept, and its standard deviation (SD). $|Q \mapsto C|_{min}$ and $|Q \mapsto C|_{max}$: Mininum and maximum number of queries mapped to a given target concept, respectively.

Then, the recurrent layer processes the vector $\mathbf{x}_i$ of each word in the social media message sequentially and produces a hidden state output $\mathbf{h}_i \in \mathbb{R}^k$, where $k \in \mathbb{Z}$ and $k > 0$. Importantly, when processing each input vector $\mathbf{x}_i$, the hidden state output $\mathbf{h}_{i-1}$ from the previous word is also recursively taken into account:

$$\mathbf{h}_i = f\left(\mathbf{h}_{i-1}, \mathbf{x}_i\right) \tag{4}$$

where f is a recurrent unit, such as long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and gated recurrent unit (GRU) (Cho et al., 2014).

Finally, the hidden state output $\mathbf{h}_l$, which is the output from processing the last word of the social media message, is used as an input of the softmax for identifying the appropriate concept, in the same manner as the vector at the fully connected layer of the CNN approach in Section 3.1.

4 Experimental Setup

4.1 Datasets

To evaluate our proposed approaches, we use three different datasets (namely, *TwADR-S*, *TwADR-L* and *AskAPatient*)[4], where the task is to map a social media phrase to a relevant medical concept. In these datasets, a given social media phrase is mapped to only one medical concept. Table 2 shows statistics for the three datasets. In particular, TwADR-S is the dataset provided by Limsopatham and Collier (2015a), which contains 201 Twitter phrases and their corresponding SNOMED-CT[5] concept. The total number of target concepts is 58, while on average a medical concept can be mapped by 3.47 queries with the standard deviation of 5.63.

The TwADR-L dataset is our new dataset that we constructed from a collection of three months of tweets (between July and November 2015), downloaded using the Twitter Streaming API[6] by filtering using the name of a pre-defined set of drugs, which have been used in the literature for ADR profiling (e.g. cognitive enhancers) (Bender et al., 2007). These tweets were sampled and then annotated by undergraduate-level linguists. This collection contains 1,436 Twitter phrases that can be mapped to one of 2,220 medical concepts from the SIDER 4 database of drug profiles[7]. Note that 1,947 from the 2,220 concepts are not relevant to any of the Twitter phrases.

For the AskAPatient dataset, we extracted the gold-standard mappings of social media messages and medical concepts from the ADR annotation collection of Karimi et al. (2015). Our AskAPatient dataset contains 8,662 phrases[8], each of which can be mapped to one of the 1,036 medical concepts from SNOMED-CT and AMT (the Australian Medicines Terminology). We expect this dataset to be less difficult than TwADR-S and TwADR-L, as the nature of blog posts is less informal and ambiguous than Twitter messages.

For each of the datasets, we randomly divide it into ten equally folds, so that our approaches and the baselines would be trained on the same sets of data. We evaluate our approaches based on the accuracy performance, averaged across the ten folds. The significant difference between the performance of our approaches and the baselines is measured using the paired t-test ($p < 0.05$).

4.2 Pre-trained Word Embeddings

As our CNN (Section 3.1) and RNN (Section 3.2) approaches require word vectors as inputs, we investigate the use of two different pre-trained word embeddings. The first word embeddings (denoted, *GNews*) are the publicly available 300-dimension embeddings (vocabulary size of 3M) that were induced from 100 billion words from Google News using word2vec (Mikolov et al., 2013b)[9], which has been shown to be effective for several tasks (Baroni et al., 2014; Kim, 2014). The second word embeddings (denoted, *BMC*) induced from 854M words of medical articles downloaded from BioMed Central[10] by using the skip-gram model from word2vec (with default parameters). The BMC embeddings also have 300 dimension. For the words that do not existing in any embeddings, we use a vector of random values sampled from $[-0.25, 0.25]$.

As an alternative, we also use randomly generated embeddings (denoted, *Rand*) with 300 dimensions, where a vector representation of each word is randomly sampled from $[-0.25, 0.25]$. This allows the investigation of the effectiveness of our approaches when the semantic information from pre-built embeddings is not available.

4.3 Parameters of Our CNN and RNN Approaches

For our CNN approach, we use the filter $\mathbf{w}$ with the window size h of 3, 4 and 5, each of which with 100 feature maps, which have shown to be effective for modelling sentences in sentiment analysis (Kim, 2014). For the RNN, we use gated recurrent unit (GRU) (Cho et al., 2014) and set the size k of the output vector of each recurrent unit to 100.

In addition, for both CNN and RNN, we use rectifier linear unit (ReLU) (Nair and Hinton, 2010) as activation functions. We also apply L_2 regularisation of the weight vectors. We train the models over a mini-batch of size 50 to minimise the negative log-likelihood of correct predictions. The stochastic gradient descent with back-propagation is performed using Adadelta update rule (Zeiler, 2012). We initially set the number of epochs for training both CNN and RNN approaches to be 100, and allow the models to update the input

[4] TwADR-L and AskAPatient datasets are available on Zenodo.org (DOI:http://dx.doi.org/10.5281/zenodo.55013).

[5] `http://www.ihtsdo.org/snomed-ct.`

[6] `https://dev.twitter.com/streaming/overview`

[7] `http://sideeffects.embl.de/`

[8] From blog posts on `http://www.askapatient.com` website.

[9] `https://code.google.com/p/word2vec/`

[10] `http://www.biomedcentral.com/about/datamining`

embeddings in the sentence matrix **S**. Later, in Sections 6.2 and 6.3, we discuss the performance achieved as we vary the number of epochs used for training the models, and the performance achieves when we allow and do not allow the models to update the input embeddings, respectively.

4.4 Baselines

We consider five different baselines as follows:

1. *TF-IDF*: A traditional term matching-based approach, using the TF-IDF score.

2. *BM25*: A traditional term matching-based approach, using the BM25 score, which has shown to be effective for several text retrieval tasks (Robertson and Zaragoza, 2009)

3. *EmbSim*: The cosine similarity between the word vector representation of a social media phrase and the description of a medical concept. If the phrase (or the description) contains several words, we represent it by adding up the values of the same dimension of the embedding of each word.

4. *DNorm*: A machine learning-based approach that exploits the relationships between words (e.g. synonyms) learned from training data (Leaman and Lu, 2014). This approach achieved state-of-the-art performance for several medical concept normalisation tasks (Suominen et al., 2013; Doğan et al., 2014). Note that we customise the open-source version[11] of DNorm to enable the mapping to a specific set of the target concepts for each dataset.

5. *P-MT*: The concept normalisation approach that translates social media texts to a formal medical text before mapping to appropriate medical concepts using the cosine similarity of their embeddings (Limsopatham and Collier, 2015a). We use the variant where the top-5 translations are used to map the medical concepts by taking the ranked position into account. We calculate the cosine similarity using either the GNews or the BMC embeddings.

6. *LogisticRegression*: A variant of our proposed approaches where we concatenate embeddings of terms (padded where necessary) in each social media phrase into a fixed-size sentence vector, before using this vector as input features for a multi-class logistic regression classifier.

Another possible baseline is a word-sense disambiguation system, such as IMS (Zhong and Ng, 2010). Nevertheless, the results from our initial experiments using IMS showed that it could not perform effectively on the three datasets. This is because the performance of IMS depends heavily on the contexts (i.e. words surrounding the input phrase); however, such contexts are not available in any of the three datasets. Therefore, we do not report the performance of IMS in this paper.

Note that for the baselines that require training data (i.e. DNorm and P-MT) and our two proposed approaches, apart from the training data provides with each fold of the datasets, we also train them using the descriptions of the target medical concepts, as these data are also used by the non-supervised baselines (i.e. TF-IDF, BM25 and EmbSim).

5 Experimental Results

In this section, we compare the performance of our CNN and RNN approaches for medical concept normalisation against the six baselines, introduced in Section 4.4. Table 3 compares the performances of our proposed approaches with the baselines in terms of accuracy on the three datasets (i.e. TwADR-S, TwADR-L, AskAPatient). Overall, as expected, the accuracy performance achieved by our approaches and the baselines on the AskAPatient dataset is higher than the TwADR-L and TwADR-S. This is due to nature use of language in Twitter, which is more ambiguous and informal than blog posts. When comparing among the existing baseline approaches, we observe that *DNorm* and *P-MT* are the most effective baselines. In particular, *DNorm* outperforms the other baselines for the TwADR-S (accuracy 0.2983) and AskAPatient (accuracy 0.7339) datasets, while *P-MT* with GNews embeddings is the most effective baseline for the TwADR-L dataset (accuracy 0.3371). In addition, term matching-based approaches, i.e. *TF-IDF* (accuracy 0.1638, 0.2293 and 0.5547, respectively) and *BM25* (accuracy 0.1638, 0.2300 and 0.5546), achieve almost similar performances, which are also comparable to the performances of *EmbSim* baselines. When comparing the effectiveness of different

[11] `http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#DNorm`

Approach	Word Embeddings	Accuracy		
		TwADR-S	TwADR-L	AskAPatient
TF-IDF	-	0.1638	0.2293	0.5547
BM25	-	0.1638	0.2300	0.5546
EmbSim	GNews	0.2494	0.2326	0.5422
EmbSim	BMC	0.1348	0.2057	0.5141
DNorm	-	**0.2983**	0.3099	**0.7339**
P-MT	GNews	0.2346	**0.3371**	0.7235
P-MT	BMC	0.1248	0.3114	0.7126
LogisticRegression	GNews	0.3186	0.3409	0.7767
LogisticRegression	BMC	0.3036	0.3548	0.7752
CNN	Rand	0.3229$^{\bullet}$	0.4267$^{*\circ\bullet}$	0.8013$^{*\circ\bullet}$
CNN	GNews	**0.4174**$^{*\circ\bullet}$	**0.4478**$^{*\circ\bullet}$	**0.8141**$^{*\circ\bullet}$
CNN	BMC	0.3921$^{*\circ\bullet}$	0.4415$^{*\circ\bullet}$	0.8139$^{*\circ\bullet}$
RNN	Rand	0.2936$^{\bullet}$	0.3791$^{*\circ\bullet}$	0.7991$^{*\circ\bullet}$
RNN	GNews	**0.3529**$^{*\circ\bullet}$	**0.3882**$^{*\circ\bullet}$	**0.7998**$^{*\circ\bullet}$
RNN	BMC	0.3331$^{\bullet}$	0.3847$^{*\circ\bullet}$	0.7996$^{*\circ\bullet}$

Table 3: The accuracy performance of our proposed approaches and the baselines. Significant differences ($p < 0.05$, paired t-test) compared to the *DNorm*, *P-MT with GNews embeddings*, and *P-MT with BMC embeddings*, are denoted *, $^\circ$ and $^\bullet$, respectively.

pre-trained embeddings used in *EmbSim* and *P-MT*, we observe that GNews is more effective than BMC for both approaches, across all of the three datasets.

Next, we discuss the performance of our *CNN* and *RNN* approaches. From Table 3, we observe that both *CNN* and *RNN* markedly outperform all of the existing baselines for all of the three datasets. When compared with *DNorm* and *P-MT with GNews* baselines, which are the most effective existing baselines, we observe that both *CNN* and *RNN* significantly ($p < 0.05$, paired t-test) outperform the two baselines for all of the three datasets. Indeed, for the TwADR-L dataset, *CNN with GNews* (accuracy 0.4478) outperforms *DNorm* (accuracy 0.3099) by 44%. In addition, the choice of embeddings has a marked impact on the achieved performance. In particular, the GNews embeddings benefit both *CNN* and *RNN* more than the BMC embeddings, which is in line with the previous finding that GNews is more useful than BMC for the *EmbSim* and *P-MT* baselines. On the other than, the randomly generated embeddings (i.e. Rand) are less useful. These results show that the semantics captured in word embeddings are useful for both *CNN* and *RNN* approaches for medical concept normalisation. However, for both *CNN* and *RNN*, the choice of embeddings that are employed has less impact on the performance for the AskAPatient dataset, which has greater number of training data.

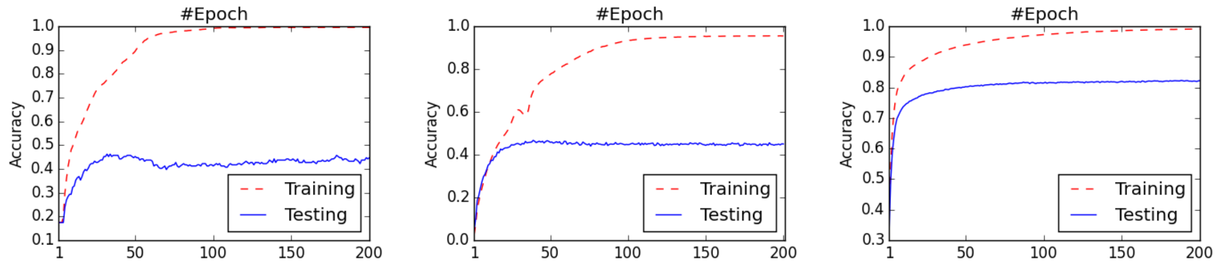
Furthermore, we observe that the *LogisticRegression* baseline, a variant of our proposed approach that uses the multi-class logistic regression instead of neural networks for identifying relevance concepts, also outperforms the all of the existing baselines. However, it performs worse than both *CNN* and *RNN* approaches. This shows that while logistic regression can exploit the semantics of embeddings of individual terms in social media texts (at the word level), it cannot learn the semantics of the whole phrase as effectively as CNN and RNN.

6 Analysis & Discussions

In this section, we further analyse the performance achieved by our proposed approaches. As the performance achieved by our CNN approach is better than that of our RNN approach, we discuss only our CNN approach in this section.

6.1 Failure Analysis

We first discuss the results achieved by the baselines and our CNN approach. As expected, we observe that all approaches perform very well for the social media phrases that lexically match with the definition of the medical concepts, e.g. the social media phrase "attention deficit disorder" is



(a) TwADR-S

(b) TwADR-L

(c) AskAPatient

Figure 3: The accuracy performance achieved by training with different numbers of epochs for the three datasets.

mapped to the medical concept 'Attention Deficit Disorder'. However, for a more complex phrases, such as "appetite on 10", "my appetite way up", "suppressed appetite", the baselines, including DNorm and P-MT, cannot effectively incorporate the modifiers of the word "appetite" in different phrases. For example, "appetite on 10", "my appetite way up" should be mapped to 'Increased Appetite', while "suppressed appetite" should be mapped to 'Loss of Appetite'. On the other hand, for social media phrases that do not have any terms in common with the definition of any medical concepts, all of the baselines performs poorly for most of the cases. For instance, even though DNorm can learn that the term "focusing" has some relationship with "concentration", it maps any phrases containing "focusing" to the 'Attention Concentration Difficulty' concept, including phrases, such as "focusing monster", which should be mapped to 'Consciousness Abnormal'. Our CNN approach could deal with most of these cases effectively, as it considers the semantic representation of the whole phrase during normalisation.

6.2 Impact of Number of Training Epochs

Next, we discuss the normalisation performance as we vary, between 1 and 200, the number of epochs used for training our CNN model. Figures 3(a), 3(b) and 3(c) show the performance in terms of accuracy achieved during training and testing for the TwADR-S, TWADR-L and AskAPatient datasets, respectively. We observe that training can be effectively achieved at around 60 - 70 epochs for the TwADR-S and TwADR-L datasets, and around 40 epochs for the AskAPatient dataset, before the performance becomes stable. We notice a gap between the performance achieved during training and testing, especially for the TwADR-S and TwADR-L datasets; however, this gap should be narrower if more training data are available.

Dataset	Accuracy	
	CNN with updated emb.	CNN with fixed emb.
TwADR-S	0.4174	**0.4369**
TwADR-L	0.4478	**0.4590**
AskAPatient	**0.8141**$^\bullet$	0.7869

Table 4: The accuracy performance of our CNN approach with the GNews embeddings, when allowing (*updated emb.*) and not allowing (*fixed emb.*) the model to update the input word embeddings. Significant difference ($p < 0.05$, paired t-test) between the performance achieved by the two variants, on each dataset, is denoted $^\bullet$.

6.3 Impact of Fixed Embeddings

In this section, we compare the performance of our CNN with GNews embeddings when we allow (*updated emb.*) and when we do not allow (*fixed emb.*) the input embeddings to be updated. Table 4 reports the accuracy performance of the two variants for the three datasets. We observe that for TwADR-S and TwADR-L datasets, which are smaller datasets (dataset size of 201 and 1,436, respectively), a better performance can be achieved if the model is not allowed to update the embeddings of the input phrases. In contrast, for the AskAPatient dataset (dataset size of 8,662), allowing the model to update the embeddings results in a significantly (paired t-test, $p < 0.05$) better performance. We observe the same trends of performance when using BMC embeddings. These results suggest that for small datasets, we should leverage semantics from pre-built word embeddings and do not allow the model to update the embeddings. Meanwhile, for a larger dataset, further performance improvement can be achieved by allowing the model to update the embeddings.

7 Conclusions

We have motivated the importance of semantics when normalising medical concepts in social media messages. In particular, as social media messages are typically ambiguous, we argue that effective concept normalisation should deal with them at the semantic level. To do so, we introduced two neural network-based approaches for medical concept normalisation, which are based on convolutional and recurrent neural network architectures. Our experimental results evaluated on three different social media datasets showed that both of our approaches markedly and significantly outperformed several strong baselines, including an existing approach that achieved state-of-the-art performance on several medical concept normalisation tasks. From the analysis of the results, we found that while some existing approaches can capture synonyms of words, they could not leverage the semantic meaning of the social media message. Our approaches overcomes this by learning the semantic representation of the social media message before passing it to a classifier to match an appropriate concept.

Acknowledgements

The authors wish to thank funding support from the EPSRC (grant number EP/M005089/1).

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA*, pages 17–21.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *IJCNLP*, pages 356–364.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Andreas Bender, Josef Scheiber, Meir Glick, John W Davies, Kamal Azzaoui, Jacques Hamon, Laszlo Urban, Steven Whitebread, and Jeremy L Jenkins. 2007. Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. *ChemMedChem*, 2(6):861–873.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *ACL*, pages 655–665.

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.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

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

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

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *ACL*, pages 302–308.

Nut Limsopatham and Nigel Collier. 2015a. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *EMNLP*, pages 1675–1680.

Nut Limsopatham and Nigel Collier. 2015b. Towards the semantic interpretation of personal health messages from social media. In *Proceedings of the ACM First International Workshop on Understanding the City with Urban Informatics*, UCUI '15, pages 27–30, New York, NY, USA. ACM.

Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard TH Tsai, Hong-Jie Dai, Naoaki Okazaki, et al. 2011. The gene normalization task in biocreative iii. *BMC bioinformatics*, 12(Suppl 8):S2.

Andrew McCallum, Kedar Bellare, and Fernando Pereira. 2012. A conditional random field for discriminatively-trained finite-state string edit distance. *arXiv preprint arXiv:1207.1406*.

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2015. Concept extraction to identify adverse drug reactions in medical forums: A comparison of algorithms. *arXiv preprint arXiv:1504.06936*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814.

Karen O'Connor, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen L Smith, and Graciela Gonzalez. 2014. Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA*, volume 2014, pages 924–933.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *arXiv preprint arXiv:1404.5367*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Eric Sven Ristad and Peter N Yianilos. 1998. Learning string-edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(5):522–532.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, pages 801–809.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 212–231. Springer.

Yoshimasa Tsuruoka, John McNaught, Sophia Ananiadou, et al. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768–2774.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*, pages 384–394.

Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *ACL*, pages 78–83.