Using an ensemble of linear and deep learning models in the SMM4H 2017 medical concept normalisation task

Maksim Belousov[1], William Dixon[2,3], Goran Nenadic[1,2]
[1]School of Computer Science, The University of Manchester, UK;
[2]Health eResearch Centre, Farr Institute, Manchester Academic Health Science Centre, The University of Manchester, UK
[3]Arthritis Research UK Centre for Epidemiology, The University of Manchester, UK

Abstract *This paper describes a medical concept normalisation system developed for the 2nd Social Media Mining for Health Applications Shared Task 3. The proposed system contains three main stages: lexical normalisation, word vectorisation and classification. The lexical normalisation stage was aimed to correct spelling mistakes and maximise the coverage of pre-trained word embeddings utilised to generate word vectors in the following stage. We experimented with three different classification models. The multinomial logistic regression model achieved higher accuracy than the recurrent neural networks with gated recurrent unit. However, the ensemble of both classification models based on the mean rule achieved the highest accuracy of* 0.885 *on the test dataset.*

Introduction

Online health communities and social media platforms such as Twitter are often used by patients to discuss various health-related experiences including personal health conditions and adverse drug reactions. However, patients rarely use official medical terms to express their symptoms and rather use descriptive expressions that explain how they feel (e.g "kills my stomach" often refers to Abdominal pain and "feel like everything that surrounds me are circling or rolling" refers to Vertigo). Attempts to mention diseases using medical terminology often result in misspelled variations (e.g. "tackacardia" instead of Tachycardia).

The medical concept normalisation task aims to map a layman description of a medical condition to a corresponding concept in a standard medical dictionary such as MedDRA®*, the Medical Dictionary for Regulatory Activities (www.meddra.org). Formally, this task can be reduced to multi-class classification problem with extremely large number of classes (for example, MedDRA has over 20,000 preferred terms in total). Since terminologies are often organised hierarchically, concept normalisation problem sometimes can be solved as hierarchical classification problem. Still some medical concepts are similar to other concepts (e.g. "Hunger" and "Increased appetite") so it is often difficult to have a unique mapping without a wider context in which a particular disease or symptom has been described.

Traditional approaches for concept normalisation are mostly based on string matching, such as rule-based term variation mapping[1] or learning edit distances between phrases[2,3]. A recent study[4], however, has demonstrated that deep learning models, particularly convolutional (CNN) and recurrent neural networks (RNN) with pre-trained word embeddings obtained from large text corpora significantly outperform previous state-of-the-art concept normalisation models on social media data. Similarly, a combination of pre-trained generic and target domain (i.e. related to the specific task) embeddings has been shown to improve the performance of sentence classification in the medical domain[5].

The goal of the 2nd Social Media Mining for Health Applications Shared Task 3 is to identify MedDRA Preferred Term (PT) code for a given colloquial or other mention obtained from drug-related discussions on Twitter. The proposed ensemble system combines linear and deep learning models that have been trained on both generic and target domain word embeddings.

System architecture

The system architecture consists of three stages: preprocessing, word vectorisation and classification. The preprocessing stage aims addressing challenges related to noisy text and is focused on lexical normalisation (i.e. spelling correction, abbreviation expansion, slang conversion), stemming and stop words removal. During the word vectorisation stage, all words in preprocessed sentences are converted into corresponding vector-space representations that

*MedDRA® is a registered trademark of the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA).

are later utilised as features. Finally, the classification stage is aimed to predict a target concept and consists of an ensemble of multiple classifiers.

Lexical normalisation

As any other social media posts, health-related discussions also have the characteristics of informal communications such as irregular grammar, misspellings, abbreviations and slang. To this end, the lexical normalisation component aimed to reduce the noise and to maximise the effectiveness (i.e. coverage) of pre-trained word embeddings used in the following stage.

Particularly, our lexical normalisation pipeline utilises three types of external resources:

- **Vocabulary** is a set of known words, used to identify unknown (or out-of-vocabulary) words (i.e. candidates for correction). It could be a list of all English words and medical terms. However, we narrowed it down to vocabulary from a given pre-trained word embedding model.
- **Mappings** are represented as translations from one word (or word form) to another, such as abbreviated to expanded forms (e.g. "hbp" to "high blood pressure") or interjection to synonymous words or phrases (e.g. "ouchy" to "hurt"). Particularly, we used abbreviations and translations collected from the Internet & Text Slang Dictionary (noslang.com) and extended it with manually curated list of popular medical abbreviations and slang observed in the training data.
- **Language models** are used to calculate a probability score of corrected phrase candidates and pick the best candidate based on the combined ranking from all models. We have utilised three different language models: a trigram model generated from Twitter drug discussions[6], a trigram model generated on 1 million sentences from popular support groups on health-related social networking site DailyStrength (www.dailystrength.org) and a bigram model generated on medical expressions parsed from DrugInformer, a search engine for pharmaceutical products and their side effects (www.druginformer.com).

Word vectorisation

Vectorisation is a step in which all words are converted into a numeric vector representation that can be used as features to train a machine learning classification model. We utilised word2vec[7] embeddings that automatically learn hierarchical representations of words by training a recurrent neural network. In the proposed classifiers we have used the following models (two of them were trained on data from generic domains and one was trained on a target domain, namely Twitter drug discussions):

- `GoogleNews`: 300-dimensional vectors obtained from a model trained on 3 million words and phrases from Google News[8]
- `Twitter`: 400-dimensional vectors obtained from a model trained on 400 million tweets[9]
- `DrugTwitter`: 150-dimensional vectors learned from 1 million user sentences about drugs on Twitter[10]

The word vectors obtained from the pre-trained models were utilised differently depending on the classification algorithm.

Classification

To predict the most suitable medical concept corresponding to the textual description of the health condition, we have used an ensemble model that combines multiple base classifiers using the mean (averaging) rule. Namely, the final prediction is made based on the highest average value for each class derived from predicted probabilities of the base learners:

$$h_{\textit{final}}(x) = \underset{j}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} d_{t,j}(x) \tag{1}$$

In particular, we have applied this ensemble method in several places in the system to utilise multiple word embeddings in a multinomial logistic regression model and also to combine predictions of our base classifiers into the final system.

We have used three different prediction models for medical concept normalisation (which correspond to the three runs submitted for evaluation):

- **`MultiLogReg`:** Multinomial logistic regression is a model that generalise logistic regression by allowing more than two discrete outcomes that makes it is suitable for multi-class problems. In order to represent the entire phrase as a vector, the mean of a set of corresponding weighted word vectors (or zero vector for unknown words) is calculated, where word weights were calculated as inverse document frequency that shows whether the word is common or rare across all phrases. The averaging rule was applied to combine target and generic domain embeddings from three different pre-trained models (`GoogleNews`, `Twitter` and `DrugTwitter`). Particularly, the word vectors obtained from each word2vec model were used to train multiple logistic regression classifiers and their predictions were combined. The logistic regression model was trained using limited-memory BFGS optimisation[11], limited to 100 iterations.
- **`Bi-GRU`:** Recurrent neural networks (RNN) have an architecture designed to handle sequences of variable lengths and therefore have been successfully used in many natural language processing tasks. Bidirectional Gated Recurrent Unit (GRU)[12] is aimed to increase the amount of input information by performing a forward and backward pass over the sequence, where backward hidden states are calculated by feeding the input sequence in the backward order. For this model we utilised only the word vectors obtained from the `GoogleNews` model, since it was trained on the largest corpora and yielded the highest performance during the preliminary evaluation on the development set. We set number of units in the GRU layer to 70% of embedding dimension. The model was trained using AdaGrad[13] optimisation algorithm with the learning rate of 0.01. For regularisation, dropout with the rate of 0.25 was applied on the Bi-GRU output.
- **`Ensemble`:** The proposed ensemble model is aimed to utilise the predictive power of both `MultiLogReg` and `Bi-GRU` models by combining their predictions using the averaging rule shown in Equation 1.

Data

The training dataset for this task contains 6,650 phrases mapped to 472 concepts (14.09 phrases per concept in average, the most popular concept *Insomnia* contains 634 phrases, whereas 170 concepts have only single mention). The average length of phrase is 2 tokens (min: 1, max: 22). The testing dataset contains 2,500 phrases.

Results and Discussion

Table 1 shows the evaluation accuracy of the three models on the test dataset. The multinomial logistic regression model (`MultiLogReg`) trained on both generic and target embeddings outperformed the Bidirectional GRU (`Bi-GRU`) model trained only on the `GoogleNews` embeddings. However, the ensemble model yielded the highest accuracy score. This suggests that `MultiLogReg` and `Bi-GRU` learn slightly different information which leads to different predictions. The ensemble model was able to pick the correct candidate in majority of cases.

Table 1: Test accuracy (%) of proposed models

Run	Test accuracy
MultiLogReg	0.877
Bi-GRU	0.855
Ensemble	0.885

We have presented comparison of predicted concepts by different models and gold-standard labels for the test data in Table 2. For example, *"weight nightmare"* was classified as Nightmare by multinomial logistic regression model, however, despite the mention of a nightmare and lack of information about the fact that the weight was increased, two other models correctly associated it with the weight gain. In the case when the concept of formication was described

as *"feeling like there's bugs under my skin"* all systems incorrectly associated it with epidermal and dermal conditions, however, despite the "skin" mention, it is actually a neurological disorder. In other cases, when none of the systems has made the correct prediction, at least one of them associated it with a similar concept. For example, *"taste buds aren't working"* was predicted as Dysgeusia (a distortion of the sense of taste) that is associated with the correct concept (Ageusia, a complete lack of taste).

Table 2: Examples of phrases and their corresponding actual and predicted concepts

Correct concept	**Phrase**	**MultiLogReg**	**Bi-GRU**	**Ensemble**
Feeling of despair	*impending sense of doom*	Somnolence	Feeling abnormal	Feeling abnormal
Formication	*feel like there's bugs under my skin*	Pruritus	Photosens. reaction	Photosens. reaction
Weight increased	*weight nightmare*	Nightmare	***Weight increased***	***Weight increased***
Abdom. discomfort	*stomach feel weird*	Feel abnormal	***Abdom. discomfort***	***Abdom. discomfort***
Ageusia	*taste buds aren't working*	Drug ineffect.	Dysgeusia	Dysgeusia
Inj. site pain	*humira injection redness*	***Inj. site pain***	Burning sens.	Inj. site infl.
Fatigue	*aren't I tired*	Insomnia	***Fatigue***	***Fatigue***
Fatigue	*me so painfully exhausted*	***Fatigue***	Insomnia	Insomnia

* Correct predictions are ***marked***.

Conclusions

We presented an ensemble system that combines linear and deep learning models for medical concept normalisation in the context of the 2nd Social Media Mining for Health Applications Shared Task 3. The lexical normalisation was performed prior to the classification in order to reduce the noise and maximise the coverage of pre-trained word embeddings generated on generic and target domains. The multinomial logistic regression model achieved higher accuracy than the recurrent neural networks with gated recurrent unit. However, the ensemble of both classifiers based on mean rule yielded the highest performance.

References

1. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2001. p. 17.

2. McCallum A, Bellare K, Pereira F. A conditional random field for discriminatively-trained finite-state string edit distance. arXiv preprint arXiv:12071406. 2012;.

3. Ristad ES, Yianilos PN. Learning string-edit distance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;20(5):522–532.

4. Limsopatham N, Collier N. Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In: ACL (1); 2016. .

5. Limsopatham N, Collier N. Modelling the combination of generic and target domain embeddings in a convolutional neural network for sentence classification. Association for Computational Linguistics; 2016. .

6. Sarker A, Gonzalez G. A corpus for mining drug-related knowledge from Twitter chatter: language models and their utilities. Data in brief. 2017;10:122–131.

7. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013;.

8. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

9. Godin F, Vandersmissen B, De Neve W, Van de Walle R. Multimedia lab@ acl w-nut ner shared task: named entity recognition for twitter microposts using distributed word representations. ACL-IJCNLP. 2015;2015:146–153.

10. Nikfarjam A, Sarker A, OConnor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association. 2015;22(3):671–681.

11. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 1995;16(5):1190–1208.

12. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:14091259. 2014;.

13. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research. 2011;12(Jul):2121–2159.