A Deep Learning Way for Disease Name Representation and Normalization

Hongwei Liu and Yun Xu

University of Science and Technology of China, Hefei, China

Abstract. Disease name normalization aims at mapping various disease names to standardized disease vocabulary entries. Disease names have such a wide variation that dictionary lookup method couldn't get a high accuracy on this task. Dnorm is the first machine learning approach for this task. It is not robust enough due to strong dependence on training dataset. In this article, we propose a deep learning way for disease name representation and normalization. Representations of composing words can be learned from large unlabelled literature corpus. Rich semantic and syntactic properties of disease names are encoded in the representations during the process. With the new way of representations for disease names, a higher accuracy is achieved in the normalization task.

Keywords: disease name representation; disease name normalization

1 Introduction

There is a rapid growth in biological research recently which results in an exponential growth of biomedical literature [1]. Extracting useful information for better use of biological literature becomes important but difficult. During tasks in biomedical text mining, disease name recognition and normalization is a fundamental task which can be further used for tasks like disease-gene [2], disease-drug [3] relation extraction. Generally, there are two steps to find disease concepts showing in literatures. The first step is to extract mentions of disease names from literatures which can be done by tools like BANNER [4]. The next important step is the normalization of various disease mentions to standardized disease vocabulary entries. An entry in standardized disease vocabulary represents a disease concept. A disease concept is uniquely identified by a disease identifier. However, a disease concept may have multiple different disease names because a disease concept can be named by its anatomical locations, symptoms, treatment, causative agent and so on. This wide variation in disease names makes the normalization task difficult.

Dictionary lookup and pattern matching algorithms were often used in early years in the task of disease name normalization. Islamaj Doan and Lu proposed an inference method [5] for disease name normalization which was mainly based on dictionary lookup and pattern matching. A string similarity is calculated during the process. A disease concept is assigned to a disease mention if their names are very similar in spelling. But with a rich variation, some disease mentions are hardly mapped to the correct disease concepts. After that, a rule-based method was proposed by Ning Kang et al. [6]. They focused on some obvious mistakes made by dictionary lookup method and designed several rules to correct the mistakes. Since the rules were designed by human, they could only deal with mistakes in limited circumstances. Recently, Robert Leaman and Islamaj Doan implemented a disease normalization tool called Dnorm [7]. It is based on a machine learning method called pairwise learning to rank (pLTR) [8]. The method learns the semantic correlation between words from training dataset and calculates a similarity score between a disease mention and a disease concept. Disease concept with the highest similarity score will be assigned to the disease mention. There are two main shortcomings of this method: 1) It is not robust enough since it is strongly dependent on the training dataset. 2) The method ignores syntactic properties which play important roles in measuring the similarity between phrases.

In this article, we propose a deep learning way for disease name representation and normalization. Word2vec [9][10] is used to generate a distributed representation for each word of disease names. Large unlabeled literature corpus can be used to train the representation. TreeLSTM [11] is used to integrate words' representations into a representation for a disease name. Finally, a simple perceptron is used to calculate a similarity score between a disease mention and a disease concept. High robustness can be achieved since the method doesn't have strong dependence on training dataset. Also, rich semantic and syntactic properties of disease names can be captured during the process. In this paper, the details of the method will be first described in the following. The results of several experiments are shown after that. Finally, a conclusion is given to summarize the the main ideas in this paper.

2 Methods

2.1 Processing pipeline

Our processing pipeline is summarized as Fig.1. For each disease name, a distributed representation is generated using Word2Vec and TreeLSTM. A similarity score is calculated between a disease mention and each disease concept name using a simple perceptron. Finally, the disease concept whose name has the highest similarity score is assigned to the disease mention. Following is the details of the pipeline.

2.2 Word2vec and TreeLSTM for distributed representation

Usually, a disease name is a phrase composed of several words. And these words are organized based on syntax and grammar to construct the disease name. In order to fully represent a disease name, we need to utilize both the meaning of composing words and the syntactic properties of the disease name. First, a

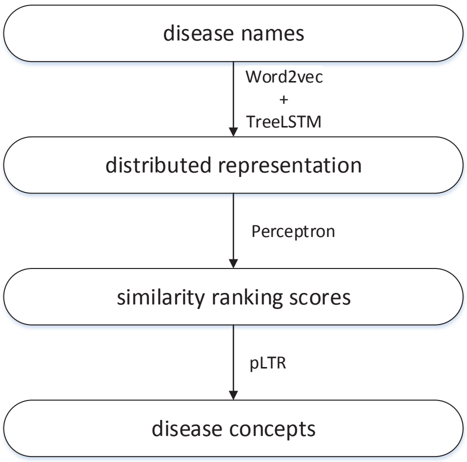

Fig. 1: Processing pipeline

distributed representation is generated for each word with the help of Word2vec. Word2Vec is a tool that learns to represent each word as a fixed-length vector in an unsupervised way. With enough training corpus, the distributed representation can reflect the meaning of words. And then, each disease name is analyzed by a dependency parser to generate a structured tree which reflects syntactic properties of that disease name. In the end, we integrate words' representations into a phrase representation based on that structured tree. Given a tree, let $C(j)$ denote the set of children of node j. Following is the concrete integrating equations from representations of children nodes to representation of their parent node:

$$\widetilde{h}_j = \sum_{k \in C(j)} h_k \tag{1}$$

$$i_j = \sigma\left(W^{(i)} x_j + U^{(i)} \widetilde{h}_j + b^{(i)}\right) \tag{2}$$

$$f_{jk} = \sigma\left(W^{(f)} x_j + U^{(f)} h_k + b^{(f)}\right) \tag{3}$$

$$o_j = \sigma\left(W^{(o)} x_j + U^{(o)} \widetilde{h}_j + b^{(o)}\right) \tag{4}$$

$$u_j = tanh\left(W^{(u)} x_j + U^{(u)} \widetilde{h}_j + b^{(u)}\right) \tag{5}$$

$$c_j = i_j \odot u_j + \sum_{k \in C(j)} f_{jk} \odot c_k \tag{6}$$

$$h_j = o_j \odot tanh\left(c_j\right) \tag{7}$$

Recursively calculating the representations of nodes from tree's left to right, from bottom to up, we can get a final representation for the tree's root. The representation of the tree's root captures both meanings of composing words and syntactic properties of the disease name so that it can be used as the representation for the disease name.

2.3 Perceptron for similarity score

For each pair $< m, n >$ where m is a disease mention extracted from corpus and n is a disease concept name in the controlled vocabulary, a score is needed to measure the similarity between them. A simple perceptron is used here which can be trained with other parts of the neural network. Distance and angle are two key points to measure the similarity of two vectors. Here, the inner product and distance are inputs to the perceptron:

$$h_p = h_L \odot h_R \tag{8}$$

$$h_s = |h_L - h_R| \tag{9}$$

$$h_d = \sigma \left(W^{(p)} h_p + W^{(s)} h_s + b^{(h)} \right) \tag{10}$$

$$\hat{p}_\theta = softmax \left(W^{(p)} h_d + b^{(p)} \right) \tag{11}$$

$$score = r^T \hat{p}_\theta \tag{12}$$

h_L and h_R are distributed representations for a disease mention and a disease concept name. The final score is a real number in interval $[0, 1]$ where higher similarity results in a higher score. $\hat{p}_\theta$ represents the distribution of different score. r is a static vector that equals to $[0, 1]$.

2.4 PLTR for concept assignment

To simplify the problem, let M represent a set of disease mentions, D represent a set of disease identifiers and N represent a set of disease concept names. Several disease concept names $n \in N$ may correspond to one identifier $d \in D$ but one disease name $n \in N$ corresponds to only one disease identifier $d \in D$ in standardized disease vocabulary. Then we can describe the normalization task as follows: for each disease mention $m \in M$, we need to assign a unique identifier $d \in D$ by comparing disease mention m and the set of disease names N. For a disease mention m, d^+ is a notation for a right identifier while d^- is a notation for a wrong identifier. Following is the process of pLTR. Given a disease mention m and a pair of identifier (d^+, d^-) in training set, we learns to give a higher similarity score for $\langle m, d^+ \rangle$ than $\langle m, d^- \rangle$. Given a new disease mention, identifier with the highest similarity score is chosen.

2.5 Training details

There are two training tasks in training process: training for words' representations and training for similarity scores. We crawl more than 8 million abstracts from PubMed website with search keywords "disease or disorder or syndrome or deficiency or dysfunction or cancer or tumor" for words' embedding training. When training parameters in TreeLSTM and perceptron, there are more incorrect names than correct names for a disease mention. We randomly sample 200 incorrect concept names and select an incorrect concept name with the highest similarity score as n^-. Then the output score of perceptron is made to be 1 for $\langle m, n^+ \rangle$ and 0 for $\langle m, n^- \rangle$. We test different values for model's hyper parameters and optimal values among test ones are chosen based on model's performance on the development set. The length of word embedding is set to be 300; the length of hidden embedding(including TreeLSTM's internal nodes and root node) is set to be 300; we use stochastic gradient descent with a learning rate of 0.01 as the optimization algorithm; cost function is the KL-divergence between the distribution of manually assigned score and the distribution of perceptron output score; library Theano [12] is used to build the TreeLSTM and the perceptron.

3 Datasets and Results

3.1 Datasets

Medical Subject Headings (MeSH)[13] and Online Mendelian Inheritance in Man (OMIM)[14] are two main terminologies for disease concepts. In 2012, a disease lexicon, namely MEDIC [15], merged OMIM into the disease branch of MeSH which makes it a deep and broad vocabulary for disease names. And a dataset called NCBI [16] disease corpus was created with MEDIC as the lexicon to help researchers develop powerful and highly effective tools for disease name recognition and normalization task. NCBI disease corpus consists of 793 PubMed [17] abstracts which are split into three subsets. We train our model with the help of training and development subset and evaluate our model on test subset. And another dataset created for Biocreative V CDR[18] task is used to test the robustness of our method.

3.2 Results

Several methods including Lucene, cosine similarity and pLTR+weight matrix are listed as a comparison. Manually-marking mentions with MeSH or OMIM identifiers which can be seen as a gold standard are fed into normalization tools. Accuracy is calculated based on result.

With Word2vec and TreeLSTM, we get richer semantic and syntactic properties of disease names. For example, "autosomal dominant disease" is correctly mapped into "genetic disease" in our method which can't be done by others. Our system successfully learns the relationship between "autosomal" and "genetic" and gives a high similarity score when comparing those two disease phrases.

Table 1: Accuracy on gold-standard mentions (train and test on NCBI)

Method	Right Number	Total Number	Accuracy
Lucene	674	960	0.702
cosine similarity	687		0.716
pLTR + weight matrix (DNorm)	789		0.822
pLTR + TreeLSTM (Ours)	819		0.853

Also, different syntactic structures with the same meaning can be captured by our method. For example, "inherited disorder" can be correctly mapped into "Disease, Hereditary".

To evaluate the robustness of two machine learning methods, we did an experiment on another dataset(BC5CDR). This time we train both our method and Dnorm on NCBI training subset only. And we evaluate two methods on BC5DR dataset which is a totally new dataset for both methods.

Table 2: Accuracy on gold-standard mentions (train on NCBI, test on BC5CDR)

Method	Right Number	Total Number	Accuracy
pLTR + weight matrix (DNorm)	3060	4424	0.715
pLTR + TreeLSTM (Ours)	3339		0.765

Though accuracy goes down for both methods, our method shows higher robustness.

4 Conclusion

In this article, we introduce a deep learning way for the disease normalization task. A distributed representation of disease name is generated with the help of Word2vec and TreeLSTM. The similarity between a disease mention and a disease concept is measured based on that distributed representation using a simple perceptron. Compared with pLTR, higher robustness is achieved since word embedding can be learned with a large unlabeled corpus. Rich semantic and syntactic properties are captured with the distributed representation in the process. And we get better results than DNorm in disease name normalization task on different datasets.

References

1. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature[J]. Database, 2011, 2011: baq036.
2. Garcia-Albornoz M, Nielsen J. Finding directionality and gene-disease predictions in disease associations[J]. BMC systems biology, 2015, 9(1): 35.
3. Yu L, Huang J, Ma Z, et al. Inferring drug-disease associations based on known protein complexes[J]. BMC medical genomics, 2015, 8(2): S2.
4. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition[C]//Pacific symposium on biocomputing. 2008, 13: 652-663.
5. Dogan R I, Lu Z. An inference method for disease name normalization[C]//2012 AAAI Fall Symposium Series. 2012.
6. Kang N, Singh B, Afzal Z, et al. Using rule-based natural language processing to improve disease normalization in biomedical text[J]. Journal of the American Medical Informatics Association, 2013, 20(5): 876-881.
7. Leaman R, Islamaj Doan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.
8. Cao Z, Qin T, Liu T Y, et al. Learning to rank: from pairwise approach to listwise approach[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 129-136.
9. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
10. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
11. Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks[J]. arXiv preprint arXiv:1503.00075, 2015.
12. Theano homepage http://deeplearning.net/software/theano/
13. Medical Subject Headings, https://www.nlm.nih.gov/mesh
14. An Online Catalog of Human Genes and Genetic Disorders, https://www.omim.org
15. Davis A P, Wiegers T C, Rosenstein M C, et al. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database[J]. Database, 2012, 2012: bar065.
16. Doan R I, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization[J]. Journal of biomedical informatics, 2014, 47: 1-10.
17. US National Labrary of Medicine, https://www.ncbi.nlm.nih.gov/pubmed
18. Li J, Sun Y, Johnson R J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction[J]. Database, 2016, 2016: baw068.