Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 2379208, 8 pages
https://doi.org/10.1155/2018/2379208

Research Article

Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF

Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen

Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen 518055, China

Correspondence should be addressed to Buzhou Tang; tangbuzhou@gmail.com

Received 26 December 2017; Accepted 13 March 2018; Published 19 April 2018

Academic Editor: Tianyong Hao

Social media in medicine, where patients can express their personal treatment experiences by personal computers and mobile devices, usually contains plenty of useful medical information, such as adverse drug reactions (ADRs); mining this useful medical information from social media has attracted more and more attention from researchers. In this study, we propose a deep neural network (called LSTM-CRF) combining long short-term memory (LSTM) neural networks (a type of recurrent neural networks) and conditional random fields (CRFs) to recognize ADR mentions from social media in medicine and investigate the effects of three factors on ADR mention recognition. The three factors are as follows: (1) representation for continuous and discontinuous ADR mentions: two novel representations, that is, "BIOHD" and "Multilabel," are compared; (2) subject of posts: each post has a subject (i.e., drug here); and (3) external knowledge bases. Experiments conducted on a benchmark corpus, that is, CADEC, show that LSTM-CRF achieves better F-score than CRF; "Multilabel" is better in representing continuous and discontinuous ADR mentions than "BIOHD"; both subjects of comments and external knowledge bases are individually beneficial to ADR mention recognition. To the best of our knowledge, this is the first time to investigate deep neural networks to mine continuous and discontinuous ADRs from social media.

1. Introduction

With rapid growth of online health social networks, such as DailyStrength.com [1], Askapatient.com [2], MedHelp.org [3], and PatientsLikeMe.com [4], more and more patients share their personal health-related information through social media posts. This information can be utilized for public health monitoring, particularly for pharmacovigilance via mining adverse drug reactions (ADRs) using natural language processing techniques.

ADRs, noxious and unintended responses to medicinal products occurring at doses normally used in man for the prophylaxis, diagnosis or therapy of disease or for the restoration, and correction or modification of physiological function [5] are an essential part of drug safety. There usually are two kinds of mechanisms to discover ADRs: (1) some of ADRs are discovered during phase III clinical trials for drug development and (2) some are revealed during postmarketing surveillance. In the case of the postmarketing surveillance, traditionally, ADRs are identified by individual patients and their physicians using official adverse event reporting systems (AERS). In recent years, there have been some attempts at mining ADRs and drug-drug interactions [6, 7] from unstructured text in clinical records, literature, and health-related social media, and their experimental results have shown the effectiveness of mining ADRs from unstructured text. In this study, we focus on recognizing ADR mentions, including continuous and discontinuous ADR mentions, from social media according to all comments to specific drugs. For example, given a user's post "I still have pain in arms and legs with much stiffness." our goal is to extract three ADR mentions, namely, "pain in arms," "pain in. . .legs," and "stiffness," where "pain in arms" and "stiffness" are continuous ADR mentions composed of continuous words and "pain in. . .legs" is a discontinuous ADR mention composed of discontinuous words.

Although there have been a number of methods proposed for recognizing ADR mentions from social media, most of

them used traditional machine learning methods such as conditional random fields (CRFs) to recognize continuous ADR mentions. In this work, we propose a deep neural network (called LSTM-CRF), which combines long short-term memory (LSTM) neural networks (a type of recurrent neural networks, RNNs) and conditional random fields (CRFs), to recognize both continuous and discontinuous ADR mentions from social media. Compared to CRFs, the advantages of LSTM-CRF lie in that LSTM neural networks have strong expressive ability to capture long context without time-intensive feature engineering. It has shown better performance than CRFs for some sequence labeling tasks such as part-of-speech (POS) tagging and named entity recognition (NER) [8].

In order to comprehensively investigate LSTM-CRF on ADR mention recognition, we compare two novel unified representations (i.e., "BIOHD" and "Multilabel") for continuous and discontinuous ADR mentions and studied effects of post subjects and external knowledge bases. All models are evaluated on a benchmark corpus, that is, CADEC, composed of 1250 forum posts for 12 drugs taken from AskaPatient.com, where each post has been manually annotated with ADR mentions. Our results show that LSTM-CRF performs better than CRF, "Multilabel" is more suitable than "BIOHD" to represent continuous and discontinuous ADR mentions, and both subject of posts and external knowledge bases are individually beneficial to ADR mention recognition.

On the whole, the contributions of this work can be summarized as follows: (1) we introduce a deep neural network that combines LSTM neural networks and CRFs to recognize continuous and discontinuous ADR mentions at first time; (2) we compare two unified representations for continuous and discontinuous ADR mentions; (3) we investigate the effects of post subjects and knowledge bases; (4) we conduct empirical evaluation of all models on a benchmark corpus.

This paper is organized as follows: in Section 2, we survey related work; Section 3 introduces the LSTM-CRF model; and Section 4 depicts our experiments in detail; we provide discussion in Section 5 and Section 6 concludes the paper. An earlier version of the paper has been presented in The 3rd China Health Information Processing Conference (CHIP-2017).

2. Related Work

In recent years, social media has been increasingly used for medical research, especially for pharmacovigilance via mining ADRs from health-related posts. Certain quantities of studies have been proposed for ADR mention recognition from corpus construction to methods. Posts from DailyStrength.com [1], Yahoo Wellness Groups [9], Askapatient.com [2], Medications.com [10], WebMD.com [11], MedHelp.org [3], SteadyHelath.com [12], PatientsLikeMe.com [4], parenting websites [13], various disease-specific forums such as Diabetes and Cancer [14], Twitter [15], Facebook [16], and other websites or forums [17] have been collected to mine ADRs. On these data, varieties of methods have been employed to recognize ADR mentions. They may fall into three categories: lexicon-based [18–29], pattern-based [30, 31], and machine learning-based [32, 33]. The earliest work, the pioneering work of Leaman et al. in 2010 [18], utilized a lexicon-based method to recognize ADR mentions from user posts regarding six drugs from DailyStrength.com. In this work, 450 out of 3600 posts were used for system development and the remanding post for system evaluation. Although lexicon-based methods can successfully recognize ADR mentions using several extensive and available ADR resources, they cannot address challenges such as idiomatic expressions and misspelled mentions. To conquer some of them, pattern-based methods over lexicon-based methods were proposed to detect inexact-match ADR mentions. For example, Yates et al. [34] designed seven patterns to recognize ADR mentions from posts regarding five drugs for breast cancer from Askpatient.com, Drugs.com, and DrugRatingZ.com. The limitation of pattern-based methods is the need for large amounts of data to generate patterns. Recently, with some annotation data available, machine learning-based methods, such as CRFs, are becoming more and more popular with promising performances, where ADR mention recognition is considered as a sequence-labeling problem. Sarker and Gonzalez [35] made a comprehensive review of text mining techniques for ADR mining before 2015. As mentioned in this review, most state-of-the-art machine learning methods are based on CRFs with rich hand-crafted features, and only continuous ADR mentions are taken into account. In 2016, Pacific Symposium on Biocomputing (PSB) launched a shared task on mining social media to exploit natural language processing techniques for ADR extraction in tweets, where subtask 2 is to automatically extract ADR mentions in user posts. In this subtask, only continuous ADR mentions were considered, and machine learning methods based on CRFs achieved best results again [33].

Actually, discontinuous entity mentions are very common in the medical domain. As reported in [36], discontinuous disorder mentions in clinical text accounted for about 10%. In social media, discontinuous ADR mentions also usually appear. Karimi et al. [37] annotated a corpus of adverse drug events including both continuous and discontinuous ADR mentions, that is, CADEC. To recognize continuous and discontinuous ADR mentions simultaneously, Metke-Jimenez and Karimi [38] followed Tang et al.'s [36] way to represent them in a unified schema and used CRFs with baseline features, including bag-of-words, character n-grams, and word shapes. To the best of our knowledge, this is the only study that considered both continuous and discontinuous ADR mentions in the task of ADR mention recognition. There are also some other studies conducted on this corpus; however, all of them only consider continuous ADR mentions or convert every discontinuous ADR mention into one or more continuous ADR mentions. For example, Tutubalina and Nikolenko [39] proposed a method, uniting RNNs and CRFs, to recognize ADR mentions on CADEC. They excluded overlaps between spans of discontinuous ADRs by selecting the longest continuous span and combining these ADRs into a single continuous ADR.

Table 1: Examples of continuous and discontinuous ADR mentions represented by "BIOHD" and "Multilabel," respectively.

Sentence 1	I	experienced	severe	pain	in	my	left	shoulder	.			
BIOHD	O	O	O	DB	DI	O	DI	DI	O			
Multilabel	O	O	O	B	I	O	I	I	O			
Sentence 2	I	still	have	pain	in	arms	and	legs	with	much	stiffness	.
BIOHD	O	O	O	HB	HI	DB	O	DB	O	O	B	O
Multilabel	O	O	O	B	I	I	O	O	O	O	B	O
				B	I	O		I			O	

Deep learning methods have been increasingly applied to solve NLP tasks in the medical domain and achieve better performance than CRFs. In the case of ADR mention recognition, Stanovsky et al. [40] employed RNNs with word embeddings trained on a Blekko medical corpus in conjunction with entity embeddings trained on DBpedia. If an entity mention was a lexical match with one of DBpedia entities, then the entity embeddings trained on DBpedia replaced word embeddings of all words in the entity mention. Tutubalina and Nikolenko [39] utilized multilayer RNNs (LSTM and GRU) with CRFs and achieved better performance than RNNs and CRFs individually. However, no study focuses on applying deep learning methods to recognize both continuous and discontinuous ADR mentions simultaneously.

3. Methods

Before recognizing continuous and discontinuous ADR mentions, we should know how to represent them. Therefore, in this section, we introduce representation schemas for both continuous and discontinuous ADR mentions at first and then machine learning methods.

3.1. Representations. Two novel representations are adopted in our study: "BIOHD" and "Multilabel." "BIOHD" is an extension of traditionally named entity representation schema "BIO" (B-beginning of a ADR mention, I-inside of a ADR mention, O-outside of a ADR mention) by introducing two additional tags: "H," a part shared by multiple medical mentions (e.g., ADR mentions), and "D," a part of a discontinuous medical mention not shared by other mentions. "Multilabel" allows a token to be labeled with more than one tag, and each tag corresponds to the position in one mention. The number of tags a token has is determined by how many mentions it appears in. Table 1 gives us examples of continuous and discontinuous ADR mentions represented by "BIOHD" and "Multilabel," respectively. In sentence 1, there is one discontinuous ADR mention "pain in. . .left shoulder," while there are two continuous ADR mentions, "stiffness" and "pain in arms," and one discontinuous ADR mention, "pain in. . .legs," in sentence 2. The ADR mentions "pain in arms" and "pain in. . .legs" in sentence 2 share the part "pain in," For convenience, a token's multiple tags can be combined into one tag. For example, sentence 2 can be tagged with "I/O still/O have/O pain/B-B in/I-I arms/I-O and/O legs/O-I with/O much/O stiffness/B-O ./O," where each token and its tag(s) are separated by "/," and multiple tags are joined by "-." In this study, the maximum number of tags a token has is set to 4 according to the statistic results from the training corpus.

3.2. LSTM-CRF. When continuous and discontinuous ADR mentions are represented by "BIOHD" or "Multilabel," recognizing them still can be formulated as a sequence labeling problem. In this study, we use LSTM-CRF to model this problem. Figure 1 illustrates the architecture of LSTM-CRF, which is composed of three layers as follows: (1) input layer, (2) LSTM-layer, and (3) CRF layer.

The input layer takes in different types of embeddings of each token. The embeddings used in this study include word embeddings, char-level embeddings, subject-related embeddings, and knowledge-based embeddings.

The LSTM layer uses bidirectional LSTM neural networks to generate hidden context representation at each position. Given a sentence $s = w_1 w_2 \cdots w_n$, where each word w_t $(1 \le t \le n)$ is represented by x_t (i.e., concatenation of word embeddings, char-level embeddings, subject-related embeddings, and knowledge-based embeddings of word w_t), the bidirectional LSTM neural networks take a sequence of embeddings $x = x_1 x_2 \cdots x_n$ as input and output a sequence of hidden context representations $h = h_1 h_2 \cdots h_n$, where $h_t = [h_{ft}^{\mathrm{T}}, h_{bt}^{\mathrm{T}}]^{\mathrm{T}}$ $(1 \le t \le n)$ is a concatenation of the outputs of both forward and backward LSTM neural networks.

The CRF layer takes a sequence of hidden context representations $h = h_1 h_2 \cdots h_n$ as input, estimates the probabilities of label sequences (from a predefined set), and returns the one of highest probability. As shown in [41], the conditional probability of a label sequence $y = y_1 y_2 \cdots y_n$ for $h = h_1 h_2 \cdots h_n$ is computed as follows (only taking the first-order linear chain CRF as an example here):

$$p_{\lambda,\mu}(y \mid h) = \frac{\exp\left(\sum_{t=1}^{n}\left(\lambda_{y_{t-1}y_t} + \left\langle \mu_{y_t}, h_t \right\rangle\right)\right)}{\exp\left(\sum_{y' \in Y(h)} \sum_{t=1}^{n}\left(\lambda_{y'_{t-1}y'_t} + \left\langle \mu_{y'_t}, h_t \right\rangle\right)\right)}, \quad (1)$$

where $\lambda_{y_{t-1}y_t}$ is the transform score for $y_{t-1}y_t$, $\langle \mu_{y_t}, h_t \rangle$ is the emission score generating y_t given h_t, and $Y(h)$ represents all possible label sequences for h.

4. Experiments

4.1. Corpus. We use a publicly available annotated corpus called CSIRO Adverse Drug Event Corpus (CADEC) from

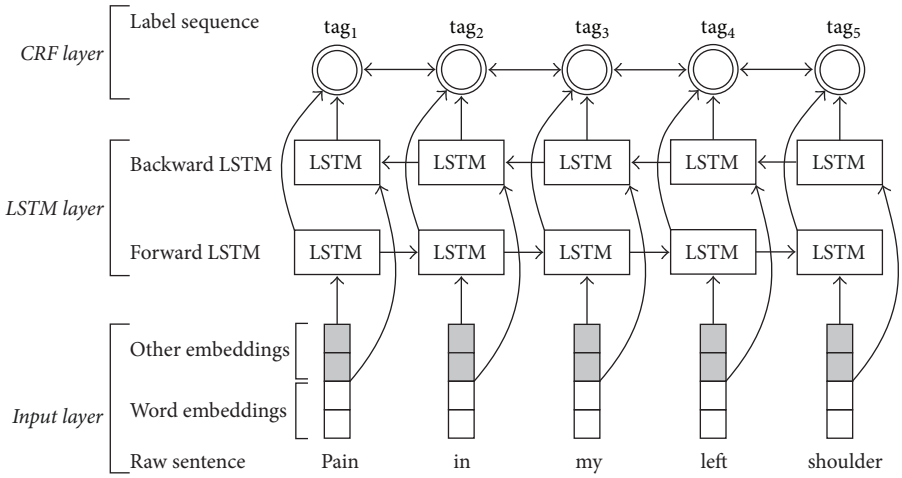

Figure 1: Architecture of LSTM-CRF for continuous and discontinuous ADR mentions recognition. "tag_i" is the label for the ith token defined in "BIOHD" or "Multilabel".

AskaPatient.com to evaluate the performance of LSTM-CRF. On AskaPatient.com, all comments are grouped by drugs. CADEC contains 1250 posts about 12 drugs {Voltaren, Cataflam, Voltaren-XR, Arthrotec, Pennsaid, Solaraze, Flector, Cambia, Zipsor, Diclofenac Sodium, Diclofenac Potassium, and Lipitor}, and the posts are manually annotated with five types of ADR-related events {ADR, Drug, Disease, Symptom, and Findings}. In CADEC, there are 6318 ADR mentions, 1000 out of which are discontinuous ADR mentions, accounting for 15.83%. Among the 1000 discontinuous ADR mentions, 918 share some parts with others, that is, overlapping.

4.2. Evaluation Metrics. We use precision (P), recall (R), F-score (F), and accuracy (Acc) to evaluate ADR mention recognition system. They are defined as follows:

$$\begin{aligned}
\text{precision} &= \frac{n_{TP}}{n_{TP} + n_{FP}}, \\
\text{recall} &= \frac{n_{TP}}{n_{TP} + n_{FN}}, \\
F\text{-score} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \\
\text{accuracy} &= \frac{n_{TP} + n_{TN}}{n_{TP} + n_{FN} + n_{FP} + n_{TN}},
\end{aligned} \tag{2}$$

where n_{TP} is the number of ADR mentions correctly predicted by a system, n_{FP} is the number of ADR mentions predicted by a system but not in the gold standard corpus, n_{FN} is the number of ADR mentions in the gold standard corpus but not predicted by a system, and n_{TN} is the number of nonentities (tagged as "O") correctly predicted.

Two kinds of criteria, that is, strict and relaxed, are adopted to calculate P, R, F, and Acc. The strict criterion refers to the fact that an ADR mention is correctly predicted only when it is exactly the same as the gold-standard one. The relaxed criterion refers to the fact that an ADR mention is correctly predicted as long as it overlaps with the gold-standard one.

4.3. Experimental Results. We reimplement Metke-Jimenez and Karimi's CRF-based system [38], select LSTM-CRF only using word embeddings and char-level embeddings as a baseline, compare LSTM-CRF with CRF, and investigate effects of ADR mention representations, post subjects, and knowledge bases on LSTM-CRF. The word embeddings are initialized by GloVe [42] and 100-dimensional pretrained embeddings on a large-scale unlabeled dataset from Wikipedia [43]. We use bidirectional LSTM neural networks to extract char-level embeddings. The LSTM neural networks take a character sequence of each word (each char is represented by character embeddings) as input and output two hidden sequence representations. The last two outputs of the bidirectional LSTM neural networks are simply concatenated into char-level embeddings. The character embeddings are randomly initialized from uniform distribution ranging in $[-1, 1]$, and its dimension is set to 25. The dimension of the char-level embeddings is also set to 25. The subject-related embeddings are randomly initialized from uniform distribution ranging in $[-1, 1]$, and their dimension is set to 10. We label each token in a sentence with BIOES (B-beginning of a ADR mention, I-inside of a ADR mention, O-outside of a ADR mention, E-end of a ADR mention, and S-a single-token ADR mention) through knowledge-based looking up, and utilize 10-dimensional embeddings, randomly initialized from uniform distribution ranging in $[-1, 1]$, to represent each token's label. The SIDER database (http://sideeffects.embl.de/) and ADR lexicon (http://diego.asu.edu/downloads/publications/ADRMine/ADR_lexicon.tsv) are two knowledge bases used in this study. All embeddings are fine-tuned during training.

As there is no fixed way to divide CADEC into two parts, one for system development and the other one for system evaluation, we adopt 10-fold cross-validation. Following the previous study [8], we set other parameters of LSTM-CRF as follows: dimension of LSTM hidden states: 100, optimizer: SGD, learning rate: 0.005, dropout rate: 0.5, and maximum

number of epochs: 200. The results of different methods are shown in Table 2, where the highest values are highlighted in bold.

Table 2 shows that the methods using "Multilabel" outperform that using "BIOHD." For example, the strict F-score of CRF using "Multilabel" is higher than CRF using "BIOHD" by 0.76% (0.6060 versus 0.5984). Compared with CRF, LSTM-CRF only using word embeddings and char-level embeddings (denoted as LSTM-CRF (baseline)) achieves better performance. When continuous and discontinuous ADR mentions are represented by "BIOHD," LSTM-CRF achieves higher strict F-score than CRF by 4.92% (0.6476 versus 0.5984), while when continuous and discontinuous ADR mentions are represented by "Multilabel," LSTM-CRF achieves higher strict F-score than CRF by 4.99% (0.6559 versus 0.6060). Both subject-based embeddings (denoted by "subject" in Table 2) and knowledge-based embeddings (denoted by "knowledge" in Table 2) are individually beneficial to LSTM-CRF. When subject-based embeddings are added, the strict F-score of LSTM-CRF using "Multilabel" is improved from 0.6559 to 0.6636, which is the highest F-score. When knowledge-based embeddings are added, the strict F-score of LSTM-CRF using "Multilabel" is improved from 0.6559 to 0.6593. When both of them are added, LSTM-CRF achieves a strict F-score of 0.6614, which is a slightly lower than the highest one (i.e., 0.6636). The differences between strict F-scores and relaxed F-scores of the same methods exceed 20%, indicating that exactly detecting ADR mentions' boundaries is not easy.

In addition, we also analyze the performances of different methods on continuous and discontinuous ADR mentions, respectively, as shown in Table 3, where the highest indices are highlighted in bold. LSTM-CRF using "Multilabel" representation and subject-based embeddings achieves the highest strict F-score of 69.94% for continuous ADR mentions, while LSTM-CRF using "Multilabel" representation and knowledge-based embeddings achieves the highest strict F-score of 41.87% for discontinuous ADR mentions. The difference between the two highest F-scores is near 20%. The baseline LSTM-CRF achieves much higher F-scores than CRF on continuous ADR mentions by about 5% and on discontinuous ADR mentions by about 8%. The methods using "Multilabel" almost always outperform that using "BIOHD" on continuous ADR mention recognition. The strict F-score difference between the methods using the two different representations ranges from 0.2% to 0.59%. For discontinuous ADR mentions, the methods using "Multilabel" always outperform that using "BIOHD" by 4.93% in average strict F-score, which is much larger than the strict F-score difference between the methods using the two different representations for continuous ADR mentions.

5. Discussion

In this study, we propose a deep neural network (i.e., LSTM-CRF) to recognize continuous and discontinuous ADR mentions from medical social media, compare it with CRF, and investigate the effects of different factors on the proposed method. Similar to other related tasks such as NER and POS tagging, LSTM-CRF outperforms CRF on continuous and discontinuous ADR mention recognition. The methods using "Multilabel" outperform that using "BIOHD." Both subject-based embeddings and knowledge-based embeddings are individually beneficial to ADR mention recognition, but when both of them are simultaneously added, the performance is not further improved.

The reason why the methods using "Multilabel" outperform that using "BIOHD" may lie in that "Multilabel" has better representation ability than "BIOHD," especially for discontinuous ADR mentions. In theory, "Multilabel" is perfect (with a coverage of 100%), while "BIOHD" is imperfect [36]. The coverage of "BIOHD" on CADEC is 89.36%. Because of this, the strict F-score difference between systems using "Multilabel" and "BIOHD" for discontinuous ADR mentions is much larger than that for continuous ADR mention, although the distributions of continuous and discontinuous ADR mentions also affect the performance. For example, there are four ADR mentions, "Extremely bad pains in hands," "Extremely bad pains in…arms," "Extremely bad pains in…muscles," and "Extremely bad pains in…quivering," recognized by "BIOHD" in "Extremely/HB bad/HI pains/HI in/HI hands/DB ,/O arms/DB ,/O and/O muscles/DB are/O constantly/O quivering/DB ./O"; however, in fact, there are only three ADR mentions, "Extremely bad pains in hands," "Extremely bad pains in…arms," and "muscles…quivering." Since different drugs have different ADRs, adding subject-based embeddings amounts to adding the relations between drugs and their ADRs, similar to relations in knowledge bases. It may be the reason for the improvement from subject-based embeddings and why simultaneously adding both the subject-based embeddings and knowledge-based embeddings does not bring further improvements.

As the distributions of continuous and discontinuous ADR mentions are imbalanced, it is easy to understand that the strict F-score difference of the same methods for continuous and discontinuous ADR mentions is not small. How to tackle data imbalance is a possible direction for further improvement, which will be considered in the future.

Although LSTM-CRF shows much better performance than CRF, the performance of LSTM-CRF is not very good, indicating that recognizing continuous and discontinuous ADR mentions from medical social media is still challenging. The main challenge is exactly determining all words or tokens of mentions, not some of them. The errors of LSTM may fall into the following three categories. (1) Some modifiers are missing. For example, there are three continuous ADR mentions, "long time flatulence," "Achilles tendon tightness," and "dizziness," in post "long time flatulence, Achilles tendon tightness, and dizziness." The first one is wrongly recognized as "flatulence." (2) Some discontinuous ADR mentions are wrongly recognized as continuous mentions by combining words or tokens between all parts. For example, the discontinuous ADR mention "hair…thinning" in sentence "I took this drug a few years ago and went off it because my hair started thinning." is wrongly recognized as a continuous ADR mention "hair started thinning." (3) There are some combination errors between continuous ADR mentions and

Table 2: Results of LSTM-CRF and CRF on CADEC.

Method	Representation	Strict				Relaxed			
		P	*R*	*F*	Acc	*P*	*R*	*F*	Acc
CRF	BIOHD	0.6707 ± 0.035	0.5402 ± 0.026	0.5984 ± 0.029	0.7037 ± 0.017	0.9175 ± 0.025	0.7700 ± 0.019	0.8373 ± 0.019	0.8892 ± 0.011
	Multilabel	0.6865 ± 0.038	0.5424 ± 0.028	0.6060 ± 0.032	0.7101 ± 0.019	0.9177 ± 0.025	0.7702 ± 0.018	0.8375 ± 0.017	0.8848 ± 0.009
LSTM-CRF (baseline)	BIOHD	0.6496 ± 0.045	0.6456 ± 0.034	0.6476 ± 0.033	0.7398 ± 0.02	0.8796 ± 0.025	**0.9213** ± 0.02	0.9000 ± 0.017	0.9346 ± 0.014
	Multilabel	0.6616 ± 0.037	0.6502 ± 0.035	0.6559 ± 0.034	0.7467 ± 0.02	0.8880 ± 0.032	0.9106 ± 0.014	0.8991 ± 0.019	0.9301 ± 0.021
LSTM-CRF + subject	BIOHD	0.6571 ± 0.037	0.6540 ± 0.029	0.6556 ± 0.032	0.7451 ± 0.019	**0.8983** ± 0.024	0.9141 ± 0.019	**0.9061** ± 0.016	**0.9501** ± 0.025
	Multilabel	**0.6780** ± 0.037	0.6499 ± 0.035	**0.6636** ± 0.033	**0.7522** ± 0.019	0.8928 ± 0.033	0.9167 ± 0.014	0.9046 ± 0.019	0.9321 ± 0.015
LSTM-CRF + knowledge	BIOHD	0.6438 ± 0.046	**0.6632** ± 0.034	0.6534 ± 0.038	0.7444 ± 0.023	0.8835 ± 0.028	0.9164 ± 0.013	0.8997 ± 0.017	0.9488 ± 0.023
	Multilabel	0.6682 ± 0.039	0.6505 ± 0.031	0.6593 ± 0.033	0.7490 ± 0.018	0.8858 ± 0.04	0.9139 ± 0.01	0.8996 ± 0.019	0.9291 ± 0.021
LSTM-CRF + all	BIOHD	0.6567 ± 0.036	0.6592 ± 0.03	0.6580 ± 0.031	0.7465 ± 0.017	0.8869 ± 0.03	0.9212 ± 0.019	0.9037 ± 0.017	0.9424 ± 0.015
	Multilabel	0.6633 ± 0.041	0.6594 ± 0.027	0.6614 ± 0.032	0.7514 ± 0.019	0.8896 ± 0.03	**0.9213** ± 0.01	0.9052 ± 0.018	0.9337 ± 0.011

Table 3: Performances of different methods on continuous and discontinuous ADR mentions, respectively (using strict criterion).

Method	Representation	Continuous ADR mention			Discontinuous ADR mention		
		P	*R*	*F*	*P*	*R*	*F*
CRF	BIOHD	0.6882	0.5950	0.6382	0.3093	0.2495	0.2762
	Multilabel	0.6896	0.5993	0.6413	0.5217	0.2405	0.3293
LSTM-CRF (baseline)	BIOHD	0.6720	0.7002	0.6858	0.3579	0.3563	0.3571
	Multilabel	0.6703	0.7062	0.6877	0.4825	0.3533	0.4079
LSTM-CRF + subject	BIOHD	0.6800	0.7092	0.6943	0.3573	0.3613	0.3593
	Multilabel	**0.6882**	0.7109	**0.6994**	0.4400	0.3263	0.3747
LSTM-CRF + knowledge	BIOHD	0.6698	0.7165	0.6924	0.3417	**0.3802**	0.3600
	Multilabel	0.6753	0.7062	0.6904	**0.5096**	0.3553	**0.4187**
LSTM-CRF + all	BIOHD	0.6813	**0.7173**	0.6988	0.3406	0.3513	0.3459
	Multilabel	0.6718	0.7154	0.6929	0.4841	0.3623	0.4144

discontinuous ADR mentions. For example, there are three ADR mentions, "Severe pain in buttocks," "Severe pain in. . .left leg," and "sciatica like symptoms," in "Severe pain in buttocks and left leg sciatica like symptoms." The last two mentions are wrongly recognized as "left leg sciatica like symptoms." Some of these errors may be corrected by using structures of sentences. It is another case of our future work. The proposed representations actually provide two ways to connect different parts of discontinues entities; therefore, the proposed methods may have potential use for relation extraction, such as drug-drug interaction extraction.

6. Conclusions

In this paper, we investigate deep neural network-based ADR mention recognition. A deep neural network (called LSTM-CRF) combining long short-term memory (LSTM) neural networks (a type of recurrent neural networks) and conditional random fields (CRFs) is proposed to recognize continuous and discontinuous ADR mentions from social media in medicine and analyze effects of ADR mention representations, subject-based embeddings, and knowledge-based embeddings. Experiments conducted on a benchmark corpus show that (1) LSTM-CRF outperforms CRF; (2) "Multilabel" representation is more suitable for continuous and discontinuous ADR mention recognition than "BIOHD"; (3) both subject-based embeddings and knowledge-based embeddings are individually beneficial for continuous and discontinuous ADR mention recognition. Moreover, some possible directions for further improvement are also presented.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper is supported in part by the following grants: National 863 Program of China (2015AA015405), National Natural Science Foundation of China (NSFC) (61573118, 61402128, 61473101, and 61472428), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20160531192358466 and JCYJ20170307150528934), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052), and CCF-Tencent Open Research Fund (RAGR20160102).

References

[1] "Online Support Groups and Forums at DailyStrength," http://www.dailystrength.org.

[2] "Ask a Patients," http://www.askapatient.com.

[3] "MedHelp Medical Support Communities," http://www.medhelp.org/forums/list.

[4] "PatientsLikeMe: live better, together," http://www.patientslikeme.com.

[5] Guideline ICHHT, "Guideline for good clinical practice," *Journal of Postgraduate Medicine*, vol. 47, no. 1, pp. 45–50, 2001.

[6] I. Segura-Bedmar, P. Martínez, and M. H. Zazo, "Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," in *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Second Joint Conference on Lexical and Computational Semantics (∗SEM)*, vol. 2, pp. 341–350, 2013.

[7] I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros, "The 1st DDIExtraction-2011 challenge task: extraction of drug-drug interactions from biomedical texts," in *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*, pp. 1–9, September 2011.

[8] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *Computation and Language*, 2015.

[9] "Yahoo! Groups," https://groups.yahoo.com/neo.

[10] "The premier community to talk about health," http://www.medications.com/.

[11] WebMD, "Better information. Better health," https://www.webmd.com/.

[12] "SteadyHealth – ask, share, contribute," http://www.steadyhealth.com/.

[13] "Parenting.co.uk - How You Can Be a Better Parent," http://www.parenting.co.uk/.

[14] "Diabetes and Cancer," http://www.diabetes.co.uk/diabetes-complications/diabetes-and-cancer.html.

[15] "Twitter. It's what's happening," https://twitter.com/.

[16] "Facebook," https://www.facebook.com/.

[17] Y. Zeng, X. Liu, Y. Wang et al., "Recommending education materials for diabetic questions using information retrieval approaches," *Journal of Medical Internet Research*, vol. 19, no. 10, 2017.

[18] R. Leaman, L. Wojtulewicz, R. Sullivan et al., "owards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks," in *Proceedings of the 2010 workshop on biomedical natural language processing*, pp. 117–125, Association for Computational Linguistics, 2010.

[19] A. Benton, L. Ungar, S. Hill et al., "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation," *Journal of Biomedical Informatics*, vol. 44, no. 6, pp. 989–996, 2011.

[20] J. Hadzi-Puric and J. Grmusa, "Automatic drug adverse reaction discovery from parenting websites using disproportionality methods," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 792–797, IEEE Computer Society, 2012.

[21] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, "Social media mining for drug safety signal detection," in *ProcEedings of the 2012 International Workshop on Smart Health and Wellbeing*, pp. 33–40, 2012.

[22] X. Liu and H. Chen, "AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums," in *International Conference on Smart Health*, pp. 134–150, Springer, Berlin, Germany.

[23] S. Yeleswarapu, A. Rao, T. Joseph, V. G. Saipradeep, and R. Srinivasan, "A pipeline to extract drug-adverse event pairs from multiple data sources," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, article no. 13, 2014.

[24] C. C. Freifeld, J. S. Brownstein, C. M. Menone et al., "Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter," *Drug Safety*, vol. 37, no. 5, pp. 343–350, 2014.

[25] I. Segura-Bedmar, R. Revert, and P. Martínez, "Detecting drugs and adverse events from Spanish health social media streams," in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pp. 106–115, 2014.

[26] X. Liu, J. Liu, and H. Chen, "Identifying adverse drug events from health social media: A case study on heart disease discussion forums," in *International Conference on Smart Health*, pp. 25–36, Springer, Cham, Switzerland, 2014.

[27] K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez, "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions," in *AMIA Annual Symposium Proceedings*, vol. 2014, artricle 924, American Medical Informatics Association, 2014.

[28] C. C. Yang, H. Yang, and L. Jiang, "Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media," *ACM Transactions on Management Information Systems (TMIS)*, vol. 5, no. 1, article no. 2, 2014.

[29] H. Sampathkumar, X.-W. Chen, and B. Luo, "Mining adverse drug reactions from online healthcare forums using Hidden Markov Model," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, article 91, 2014.

[30] A. Nikfarjam and G. H. Gonzalez, "Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments," in *AMIA Annual Symposium Proceedings*, vol. 2011, article 1019, American Medical Informatics Association, 2011.

[31] A. Yates and N. Goharian, "ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites," in *European Conference on Information Retrieval*, pp. 816–819, Springer, Berlin, Germany, 2013.

[32] K. Jiang and Y. Zheng, "Mining Twitter data for potential drug effects," in *International Conference on Advanced Data Mining and Applications*, pp. 434–443, Springer, Berlin, Germany, 2013.

[33] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association* , vol. 22, no. 3, pp. 671–681, 2015.

[34] A. Yates, N. Goharian, and O. Frieder, "Extracting adverse drug reactions from social media," in *AAAI Conference on Artificial Intelligence*, pp. 2460–2467, 2015.

[35] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *Journal of Biomedical Informatics*, vol. 53, pp. 196–207, 2015.

[36] B. Tang, Q. Chen, X. Wang et al., "Recognizing disjoint clinical concepts in clinical text using machine learning-based methods," in *AMIA Annual Symposium Proceedings*, vol. 2015, article 1184, American Medical Informatics Association, 2015.

[37] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "Cadec: A corpus of adverse drug event annotations," *Journal of Biomedical Informatics*, vol. 55, pp. 73–81, 2015.

[38] A. Metke-Jimenez and S. Karimi, "Concept extraction to identify adverse drug reactions in medical forums: a comparison of algorithms," *Artificial Intelligence*, 2015.

[39] E. Tutubalina and S. Nikolenko, "Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews," *Journal of Healthcare Engineering*, vol. 2017, Article ID 9451342, 9 pages, 2017.

[40] G. Stanovsky, D. Gruhl, and P. N. Mendes, "Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, pp. 142–151, 2017.

[41] J. Lafferty, A. McCallum, and F. C. N. Pereira, "CondiTional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[42] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 14, pp. 1532–1543, 2014.

[43] "Wikipedia," https://www.wikipedia.org/.