Drug Safety (2019) 42:135–146
https://doi.org/10.1007/s40264-018-0764-x

ORIGINAL RESEARCH ARTICLE

Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks

Bharath Dandala[1] · **Venkata Joopudi**[1] · **Murthy Devarakonda**[1,2]

Published online: 16 January 2019
© Springer Nature Switzerland AG 2019

Abstract

Background and Significance Adverse drug events (ADEs) occur in approximately 2–5% of hospitalized patients, often resulting in poor outcomes or even death. Extraction of ADEs from clinical narratives can accelerate and automate pharmacovigilance. Using state-of-the-art deep-learning neural networks to jointly model concept and relation extraction, we achieved the highest integrated task score in the 2018 Medication and Adverse Drug Event (MADE) 1.0 challenge.

Methods We used a combined bidirectional long short-term memory (BiLSTM) and conditional random fields (CRF) neural network to detect medical entities relevant to ADEs and a combined BiLSTM and attention network to determine relations, including the adverse drug reaction relation between medication and sign or symptom entities. Using these models, we conducted three experiments: (1) separate and sequential modeling of entities and relations; (2) joint modeling where relations between medications and sign or symptoms determined ADE and indication entities; (3) use of information from external resources such as the US FDA's adverse event database as additional input to the second method.

Results Joint modeling improved the overall task accuracy from 0.62 to 0.65 *F* measure, and the additional use of external resources improved the accuracy to 0.66 *F* measure. Given the gold-standard medical entity labels, the joint model plus external resources method achieved *F* measures of 0.83 for ADE-relevant medical entity detection and 0.87 for relation detection.

Conclusion It is important to use joint modeling techniques and external resources for effectively detecting ADEs from clinical narratives in electronic health record (EHR) systems. While the extraction of entities and relations individually achieved high accuracy, the integrated task still has room for further improvement.

Key Points

Harmful side effects of medications are an important concern because of their economic and health impact.

Physician-authored clinical narratives are a reliable source for identifying such side effects, but the technical challenges of automatically analyzing them remains a limiting factor.

This study demonstrates that recent advances in neural network-based deep-learning techniques provide an effective means to address this limitation.

Part of a theme issue on "NLP Challenge for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0)" guest edited by Feifan Liu, Abhyuday Jagannatha and Hong Yu.

✉ Murthy Devarakonda
mvd@acm.org

[1] IBM Research, Yorktown Heights, NY, USA

[2] Biomedical Informatics, Arizona State University, Tempe, USA

1 Introduction

An adverse drug event (ADE) is commonly defined as an injury resulting from medical intervention related to a drug. Prevention, early detection and mitigation of ADEs improve patient safety. Consequently, reducing preventable patient harm is emphasized by national, regional and global health authorities. Electronic health records (EHRs) contain provider-recorded documentation of ADEs in clinical narratives and are an important source for pharmacovigilance. Natural language processing (NLP)-based extraction of ADEs from the clinical narratives in EHRs can simplify and automate pharmacovigilance.

Effective NLP techniques for medical entity and relation identification are a fundamental requirement in automatic ADE extraction. Accuracy of these foundation analytics will significantly impact ADE curation and pharmacovigilance. Combined bidirectional long short-term memory (BiLSTM) and conditional random fields (CRF) models [1] have previously been shown to accurately recognize entities in biomedical and clinical corpora [2–5].

Therefore, we studied the use of BiLSTM-CRF for recognizing medical entities, formally known as named entity recognition (NER), related to ADEs in clinical narratives.

Attention mechanism, introduced in Bahdanau et al. [6], is a technique often used in neural translation of text. The attention mechanism allows neural networks to selectively focus on specific information, which has benefited several NLP tasks such as factoid question answering [7], machine translation [6] and relation classification [8]. In this study, we used BiLSTM with attention mechanism for classifying ADE (and other) relations in clinical narratives.

This research was motivated by the 2018 Medication and Adverse Drug Event (MADE) 1.0 challenge [9], which consisted of three tasks:

1. Detect mentions of medication name and its attributes (dosage, frequency, route, duration), as well as mentions of ADEs, indications, other signs or symptoms (SSLIF) and severity.
2. Given the gold standard entity annotations, identify the attributes of a medication, relations between medications and ADEs (called "adverse" relations), medications and indications (called "reason" relations), and severity of an ADE or sign or symptom.
3. An integrated system of the two tasks, where entities recognized by the system in task 1 (in place of gold-standard annotations) are used for relation identification.

Figure 1 illustrates the key tasks and shows a few synthesized sentences (based on the original sentences) from a clinical note with entities and relations that need to be extracted. Note that the relations may exist between entities anywhere in the note, spanning across multiple sentences.

We studied three methods using the neural networks for the adverse reaction extraction:

1. *Sequential modeling* The traditional approach of sequentially extracting medical entities first and then relations among them.
2. *Joint modeling* A joint modeling approach where relations between medications and signs or symptoms were used to determine ADEs and indication entities.
3. *Joint modeling + external resources* Method 2 enhanced with information from external resources such as the US FDA's Adverse Event Reporting System (FAERS) database [10] for a medication as additional input.

Only the results of the second method were submitted to the MADE 1.0 challenge. Our system achieved first place in the integrated final task 3 of the challenge and second place in tasks 1 and 2. The accuracy analysis of the three methods showed that the joint modeling technique improved performance (F measure) by nearly 3% points (4.5% relative) over the traditional approach, and the addition of information from FAERS [10] further improved the system performance by one more percentage point (1.4% relative)—achieving an overall F measure of 0.661.

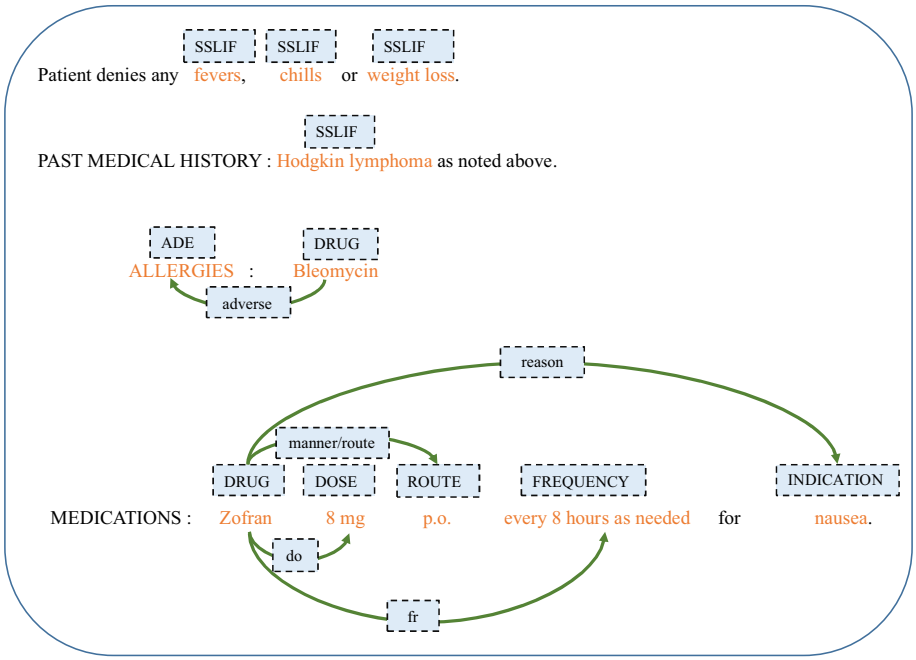

Fig. 1 Examples of the entities and relations that are a part of the adverse drug reaction extraction task

2 Methods

With recent advances in NLP research, several neural network architectures have been successfully applied to entity and relation extraction tasks. Specifically, BiLSTM-based architectures have proven to be effective [1, 8, 11, 12]. We now describe how they are used for entity and relation identification in our system.

2.1 Entity Extraction

Long short-term memory (LSTM) [13] is a type of recurrent neural network (RNN) that models interdependencies in sequential data and addresses the so-called vanishing or exploding gradients problem [14] of vanilla RNNs by using an adaptive gating mechanism. Unidirectional LSTMs do not utilize future contextual information. BiLSTM [15, 16] addresses this by using two independent LSTMs (forward and backward) in which one processes the input sequence in the forward direction and the other processes the input in the reverse direction.

Although BiLSTM networks can capture long-distance interdependencies, research suggests that additionally capturing the correlations between adjacent labels can help in sequence labeling problems [1, 17, 18]. CRF [19] helps in capturing these correlations. Therefore, similar to Huang et al. [1], we used BiLSTM-CRF for entity extraction, as shown in Fig. 2.

Given an input sequence $x = (x_1, x_2, \ldots, x_t)$, where t is the sequence length, LSTM hidden state at timestep t is computed by:

$$
\begin{aligned}
i_t &= \sigma(W^i x_t + U^i h_{t-1} + b^i) \\
f_t &= \sigma(W^f x_t + U^f h_{t-1} + b^f) \\
o_t &= \sigma(W^o x_t + U^o h_{t-1} + b^o) \\
g_t &= \tanh(W^g x_t + U^g h_{t-1} + b^g) \\
c_t &= f_t \otimes c_{t-1} + i_t \otimes g_t \\
h_t &= o_t \otimes \tan h(c_t),
\end{aligned} \tag{1}
$$

where $\sigma(\cdot)$ and $\tan h(\cdot)$ are the element-wise sigmoid and hyperbolic tangent functions, $\otimes$ is the element-wise multiplication operator, and i_t, f_t, and o_t are the input, forget, and output gates. Lastly, h_{t-1} and $c_{t-1}$are the hidden state and memory cell of previous timestep, respectively.

The forward LSTM computes the forward hidden states ($\overrightarrow{h_1}$, $\overrightarrow{h_2}$, …, $\overrightarrow{h_t}$), while the backward LSTM computes backward hidden states ($\overleftarrow{h_1}$, $\overleftarrow{h_2}$, …, $\overleftarrow{h_t}$). Then, for each timestep t, the hidden state of the BiLSTM is generated by concatenating $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ as in:

$$
h_t = (\overrightarrow{h_t}, \overleftarrow{h_t}). \tag{2}
$$

Given an observation sequence $h = [h_1, h_2, \ldots, h_t]$ (outputs from BiLSTM), CRF jointly models the probability of the entire sequence of labels $y = (y_1, y_2, \ldots, y_t)$ and we denote φ as the set of all possible label sequences. Using a linear-chain CRF model, the conditional probability of the output sequence given the input hidden state sequence can be written as:

$$
P(y|h; W, b) = \frac{\prod_{i=1}^{t} \exp\left(W^T_{y_{i-1}, y_i} h + b_{y_{i-1}, y_i}\right)}{\sum_{y' \in \varphi} \prod_{i=1}^{t} \exp\left(W^T_{y'_{i-1}, y'_i} h + b_{y'_{i-1}, y'_i}\right)},
$$

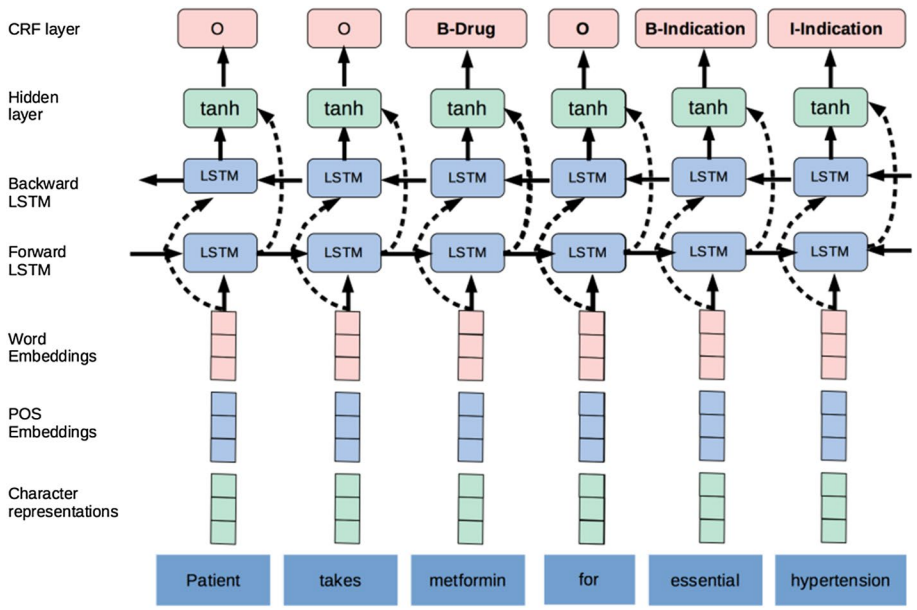

Fig. 2 Combined bidirectional long short-term memory (BiLSTM) and CRF (conditional random fields) neural network for the entity extraction. *POS* parts of speech

where W and b are weight matrices and their subscripts indicate the weight vector for the given label (y_{i-1}, y_i).We used maximum conditional likelihood estimates to train the CRF layer. For a training dataset $\{(h_i, y_i)\}$, the final log-likelihood is:

$$L(W, b) = \sum_{(h_i, y_i)} \log P(y_i|h_i; W, b).$$

For the decoding phase, a Viterbi algorithm was used to generate the optimal label sequence y^*:

$$y^* = \arg\max_{y \in \varphi} P(y|h; W, b).$$

Our neural network model used a comprehensive representation of tokens from the text as inputs. For each token, we used embeddings of its character, word level, and parser-provided syntactic elements. A convolutional neural network (CNN) [20] was used to encode character-level embedding of a word.

2.2 Relation Identification

The attention mechanism, introduced in Bahdanau et al. [6], is a technique often used in neural translation of text. It allows the networks to selectively focus on specific information. This benefited several NLP tasks such as factoid question answering [7], machine translation [6] and relation classification [8]. Here, we used the attention mechanism for the relation classification task, similar to the implementation in Zhou et al. [8], but the addition of the knowledge layer is novel (see Fig. 3).

Formally, let H be a matrix consisting of output vectors $[h_1, h_2, \ldots, h_t]$ (the outputs from the BiLSTM network), the representation r of the input is formed by a weighted sum of these output vectors:

$$\begin{aligned} M &= \tan h(H) \\ \alpha &= \text{softmax}(w^T M), \\ r &= H\alpha^T \end{aligned} \tag{3}$$

where $H \in R^{d^w \times t}$, d^w is the dimension of vectors, w^T is the transpose of the trained parameter vector. We obtain the final representation from:

$$h^* = \tan h(r). \tag{4}$$

This network takes tokens, entity types (outputs of entity extraction model) and positional indicators around source and target concepts as inputs. As mentioned earlier, this challenge requires identifying both intra- and inter-sentential relationships. Table 1 shows the number of inter- or intra-sentential relationships in training data between two entity types. In principle, the entities participating in an inter-sentential relation may occur anywhere in a document, which results in a large number of possible entity pairs that should be considered. While it may seem expedient to consider every possible entity pair in a clinical note as a potential relation, computing it will be computationally very expensive. Additionally, a very large proportion of these relations will serve as negative relation instances, resulting in a highly unbalanced dataset. Research [21, 23] suggests that a model trained over such an imbalanced dataset may not optimally differentiate among positive and negative relations. To address this, we developed a machine learning model for a priori refinement of negative instances using a set of structural and heuristic-based features. Using this model, our method generated candidate relations, which we call the "candidate relations generation phase." In this phase, for each entity pair, we extracted the following features:

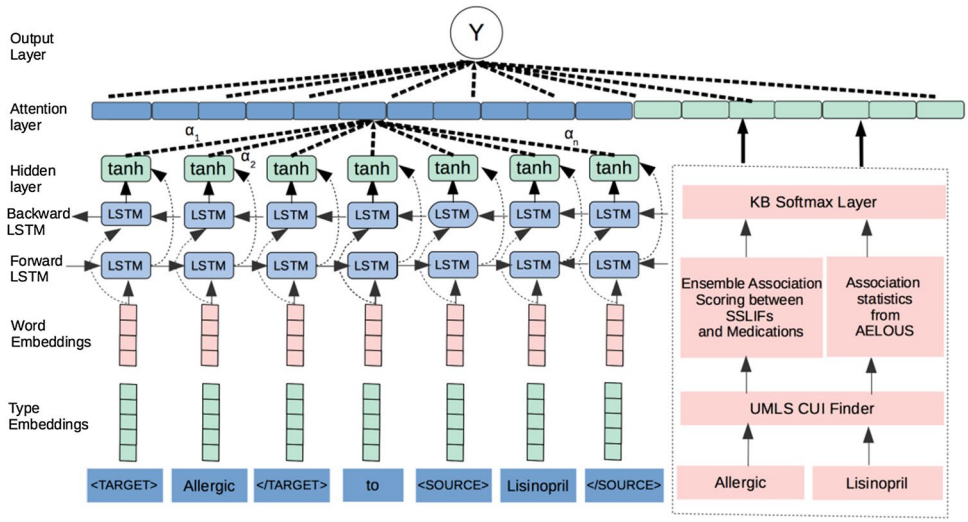

Fig. 3 Combined bidirectional long short-term memory (BiLSTM) and attention layer neural network for relation identification. The elements in the right-most box were used to add "knowledge" from external resources. *KB* knowledge base

Table 1 Number of inter- and intra-sentential relations for each relation type

Relation type	Inter-sentential relations	Intra-sentential relations
Adverse	647 (32)	1435 (68)
Reason	2307 (51)	2243 (49)
do	113 (2)	5053 (98)
fr	473 (12)	3688 (88)
Manner/route	34 (2)	2056 (98)
Severity_type	42 (1)	3424 (99)
du	79 (9)	827 (91)
Total	3695 (100)	18,726 (100)

Data are presented as *N* (%)

do dosage, *du* duration, *fr* frequency

- Since each relation type has a dominant pair of source and target entity types, we used the source and target entity types as features. For example, the majority of dosage (*do*) relation instances in the training data have *Drug* as the source entity and *Dose* as the target entity, although a handful of *Dose* to *Dose* relations were also marked with the same relation label.
- We developed a rule-based method to identify section boundaries in clinical notes (e.g., medication section, assessment and plan section, etc.), and used the names of the sections where an entity pair occurred as a feature. We also used the number of sections between the two entities as an additional feature.
- The number of sentences between the entity pair.
- The number of tokens between the entity pair.
- The count of entity types that appear between the entity pair.

We used an alternating decision tree (ADT) to train the model. We empirically determined the optimal threshold value by computing the precision-recall curve on the development dataset. At this optimal threshold value, we were able to remove 92% of candidate negative instances yet retain 98% of positive instances. However, this still left a large number of negative instances to be considered. Research on inter-sentential relation extraction [21–23] suggests addressing this issue either by under-sampling the negative class or by training a cost-sensitive classifier. During training, for each epoch, we sampled as many negative instances as the number of entity pairs with corresponding types. For example, if we had *n* positive entity pairs of type *SSLIF-Drug*, we sampled *n SSLIF-Drug* pairs from negative instances. Finally, for each pair of entities, sentences in which the entities appeared, as well as the sentences between them, served as the contextual input to the model.

2.3 Dataset

The total dataset contained 1089 de-identified clinical notes of 21 patients with cancer, of which 213 were the unseen test dataset and 876 were the training dataset. We used 86 clinical notes of the training dataset as the development set for model tuning. Each clinical note was manually annotated, identifying medications (drug name, dosage, route, frequency, duration), ADEs, indications, SSLIFs, and relations among those entities.

Table 2 shows the statistics for the entities in the training and test datasets. We observed that SSLIFs constituted the largest percentage of instances (about 50%) in the datasets, whereas drugs and ADEs were only about 20% and 2–4% of instances, respectively. An SSLIF was labeled as an ADE if the context in the clinical note implied it was a side effect of a drug; it was labeled as an indication if the context implied it was an affliction that a provider was actively treating with a drug.

Table 3 shows the relation types and their statistics in the training and test datasets. Dosage relations were the largest fraction of the relations, with about 21–22%, whereas the adverse relations accounted for only about 9–13%. These statistics indicated an imbalanced distribution of entities and relations that the methods need to consider.

2.4 Text Preprocessing

Sentence boundary detection (SBD) is a critical preprocessing task for many NLP applications. It is often treated as a solved problem and carried out using default approaches in off-the-shelf NLP toolkits. However, recent research [24] suggested that SBD remains a difficult and critical problem in the clinical domain, and renewed efforts are needed. One important challenge is that authors of clinical notes frequently indicate sentence ends by layout and not by punctuation. Thus, an SBD algorithm can sometimes incorrectly interpret physically adjacent text segments as being part of the same sentence. To address this, we used medical domain-adapted English Slot Grammar parser [25], which overcomes this problem by running a preprocessor that is sensitive to low-level features such as punctuation, capitalization, text-wrap properties, and indentation to detect implicit sentence breaks.

2.5 Experiments

Extraction of entities and relations from text has traditionally been treated as a pipeline of two separate subtasks: entity recognition and relation extraction. Thus, in our first method, called sequential modeling, we first applied our BiLSTM-CRF model introduced in Sect. 2.1 for entity recognition, a task typically addressed by assigning BIO (begin, inside,

Table 2 Entities in the dataset

Annotation	Training data		Test data		Example	Description
	Annotations	No. of distinct annotations	Annotations	No. of distinct annotations		
SSLIF	34,056 (50.2)	7243	5328 (47)	1614	Worsening renal function	All signs and symptoms
Drug	13,507 (19.9)	1231	2395 (21.1)	420	Vicodin	Name of the drug
Dose	4893 (7.2)	805	801 (7.1)	253	One tablet, tapered	Dosage of the drug
Frequency	4147 (6.1)	615	659 (5.8)	197	Daily	Frequency of the prescribed drug
Severity	3374 (5.0)	417	534 (4.7)	104	Significant, slightly	Severity of disease or symptom
Indication	3168 (4.7)	872	636 (5.6)	217	Swelling around his eye	Affliction that is being treated with a drug
Route	2278 (3.4)	108	389 (3.4)	42	Subcutaneously	Route in which the drug is given
ADE	1509 (2.2)	423	431 (3.8)	160	Vertigo	SSLIF that is a side effect of a drug
Duration	765 (1.1)	161	133 (1.2)	44	Lifelong, week	Duration of the drug
PHI	84 (0.1)	33	27 (0.2)	16	St. Vincent hospital	Unannotated PHI
Total	67,781 (100)	–	11,333 (100)	–	–	–

Data are presented as *N* (%) unless otherwise indicated

ADE adverse drug event, *PHI* protected health information, *SSLIF* other signs or symptoms

Table 3 Relations in the dataset

Relation type	No. of annotations in training data	No. of annotations in test data	Description
Do(sage)	5177 (22)	866 (21)	Relation between dosage and drug
Reason	4554 (20)	876 (21)	Drug prescribed to treat particular indication
Fr(equency)	4419 (19)	730 (18)	Relation between frequency and the drug
Severity_type	3476 (15)	559 (13)	Relation between severity and SSILF
Manner/route	2551 (11)	455 (11)	Relation between route and drug
Adverse	2082 (9)	530 (13)	Relation between adverse reaction and drug
Du(ration)	906 (4)	147 (4)	Relation between drug and duration
Total	23,165 (100)	4163 (100)	–

SSLIF other signs or symptoms

and outside) labels to each word, indicating the token's position within an entity mention as well as its type (as shown in Fig. 4). Sentences served as logical units of contextual information for the entity extraction task. Subsequently, we applied the attention-BiLSTM model to relation identification on the entity pairs that were extracted from clinical narratives and filtered as described in Sect. 2.2.

Overall, the sequential modeling method performed fairly well on categories such as medications and their associated constituents but struggled on the more challenging and important categories, "reason" and "adverse" relation types. Subsequent error analysis revealed several categories of errors. Among these, misclassifying ADEs or indications as SSLIFs was a major error category, highly critical to the overall accuracy. Further analysis revealed two distinct issues: (1) document-level contextual information was vital and (2) domain knowledge can be beneficial in identifying these clinical entity types. Consequently, we tried to address these two issues to improve identification of the reason and adverse relation types.

An important characteristic of signs and symptoms (SSLIFs, ADEs or indications) is that "the type of these entities is determined by the relationship it keeps". By definition, a certain sign or symptom is marked as an ADE or indication by its relationship to one or more medications. Furthermore, only 61% of the ADEs and 46% of indications participate in an adverse or reason relationship with a medication within the same sentence. Thus, any entity-extraction model that relies only on contextual information within a sentence is insufficient, which highlights the need for a better approach to recognizing the ADE and indication entities. To address this issue, we performed our entity extraction over two steps. In the first step, we used a BiLSTM-CRF neural network to model generic entity types. Generic entity types were obtained by replacing the ADE and indication labels with the SSLIF label in the original training data. In

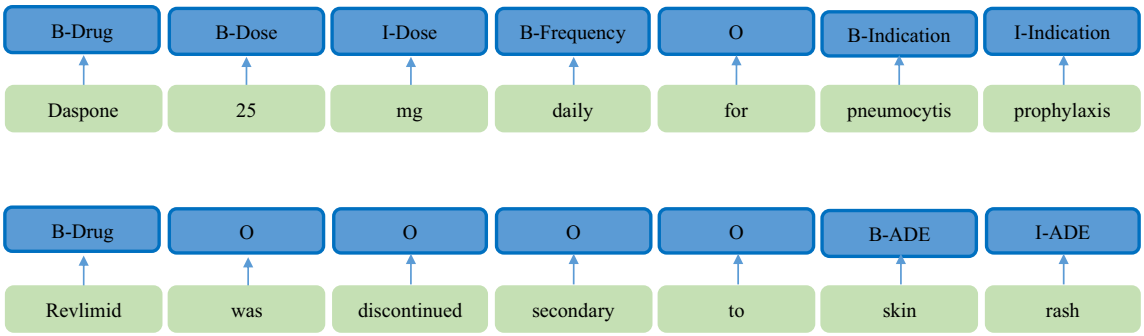

Fig. 4 BIO tagging for the entity extraction

the second step, we used the predictions from the relation identification task to infer the correct type from the generic type. Thus, for a given SSLIF, if our model predicted that it participated in an adverse or reason relationship with any medication in the clinical note, the corresponding SSLIF type was updated to ADE or indication, respectively. We called this method "joint modeling" of entities and relations.

Another important characteristic we observed is that contextual information present in the current document is not sufficient to determine adverse and reason relationships, thus indicating the importance of external knowledge. One such example is shown in Fig. 5. The third example in the figure does not contain any words that inform the relationship. However, the relationship is implicitly understood by medical experts. Effective knowledge resources have long been known to influence the effectiveness of learning algorithms [26, 27]. Therefore, we experimented with using prior medical knowledge in our relation extraction system, and we called this method "joint modeling + external resources".

Specifically, for a drug–SSLIF pair, we incorporated additional features obtained using two distinct systems, one introduced in Dandala et al. [28] and the other introduced in Banda et al. [29]. Both these systems take two sets of unified medical language system (UMLS) [30] concept unique identifiers (CUIs) as input, with one set being the CUIs for an SSLIF and the other set being CUIs for a drug. We obtained UMLS CUIs for each SSLIF and drug using an ensemble system described in Rajani et al. [31]. The system in Dandala et al. [28] returns a single score between 0 and 1 (1 being the best), indicating the strength of association. AELOUS, the system in Banda et al. [29], curates and normalizes the collaboratively captured reports in FAERS [10] and provides two scores—the proportional reporting ratio and reporting odds ratio—which we normalized to the range from 0 to 1 (1 being the best). The scores were additional inputs to the attention-BiLSTM model as shown in Fig. 3.

2.6 Experimental Settings and Metrics

We used 10% of the training data as the development dataset to tune the models and the remaining 90% of the training dataset for training the neural network models. We fixed the word embeddings size to 200, character embeddings size to 50 and part-of-speech embeddings length to 20. The part-of-speech and character embeddings were initialized with random values. Micro averaged standard precision, recall,

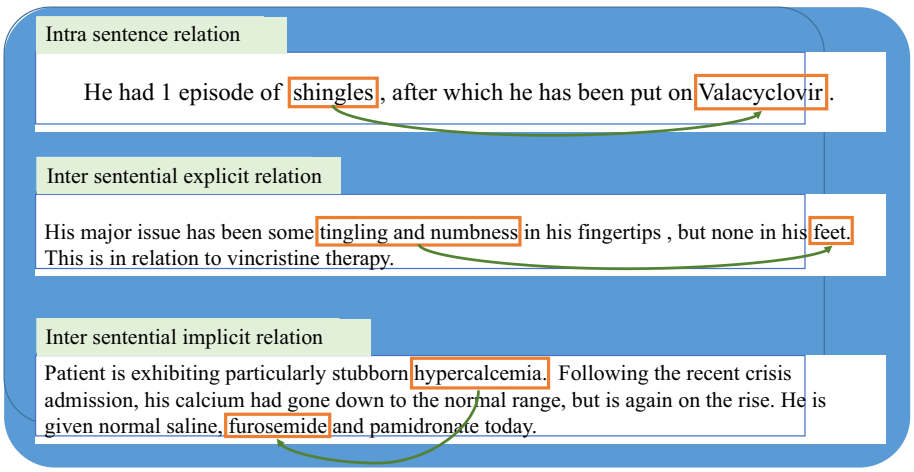

Fig. 5 Different types of relations in the dataset. Especially note that the third relation has no indicative words in the context

and F measures [58] were used as evaluation metrics for the entity extraction and relation classification tasks.

2.7 Hyperparameter Tuning

Our models include four hyperparameters: the dropout rate, learning rate, regularization parameter, and hidden layer size. The hyperparameters for our models were tuned on the development set for each task. Research has suggested that using dropout mitigates over-fitting and is especially beneficial to the NER task [11]. We experimented by tuning the hyperparameters with different settings: dropout rates (0.0, 0.1, 0.2, 0.3, 0.4 and 0.5), hidden layer sizes (100, 150, 200) and regularization parameter ($1e^{-5}$, $1e^{-6}$, $1e^{-7}$, $1e^{-8}$). We chose Adam [32] as our stochastic optimizer and tuned the learning rate at ($1e^{-2}$, $1e^{-3}$, $1e^{-4}$). We used early stopping [16] based on performance on the development dataset.

3 Results

3.1 Optimal Hyperparameter Values

We observed the best performance at around 20 epochs and 15 epochs for entity and relation extraction, respectively. We used both dropout and L2 regularization for optimizing the network parameters. Table 4 shows the neural network parameters we used after tuning.

Table 5 shows results on the challenge test dataset for all our methods. For each method, we trained 40 different models, where each of them was trained on a randomly shuffled training dataset from 1089 de-identified clinical notes. Each of these models was tested on the test dataset, and we calculated the mean precision, recall, F measure and standard deviations from the results. We also computed 95% confidence intervals of the mean F measure. Furthermore, we performed pair-wise t test for performance differences in mean F measure among the three methods. The performance differences were statistically significant at $p < 0.05$ for each pair of the methods.

Figures 6, 7 and 8 present the results for entity extraction, relation extraction with gold labels, and relation extraction with system labels (integrated task), broken down by entity or relation type for each of the three methods. We discuss the specific performance results of the three methods in the following subsections.

3.2 Sequential Modeling

The sequential modeling method achieved an F measure of 0.829 for entity extraction, 0.858 for relation classification using gold labels for entities, and 0.624 for the integrated relation extraction task. High F measure was achieved in detecting medications, its attributes, and relations between them (see Figs. 6, 7). However, performance in extracting ADE and indication concepts was poor; in particular, recall

Table 4 Neural network tuned parameters

Parameter	Sequential			Joint			Joint + external resources		
	Concept extraction	Relation classification	Relation extraction	Concept extraction	Relation classification	Relation extraction	Concept extraction	Relation classification	Relation extraction
Dropout	0.4	0.4	0.5	0.5	0.4	0.5	0.5	0.4	0.4
Learning rate	0.02	0.03	0.03	0.02	0.03	0.02	0.02	0.03	0.01
Regularization	1e−7	1e−6	1e−6	1e−5	1e−6	1e−5	1e−5	1e−6	1e−5
Hidden layer size	150	100	100	150	100	100	150	100	150

Table 5 Overall accuracy results for the three methods

Task	Sequential			Joint			Joint + external resources		
	Mean precision	Mean recall	Mean F measure	Mean precision	Mean recall	Mean F measure	Mean precision	Mean recall	Mean F measure
Concept extraction	0.847	0.812	0.829 ± 0.05	0.846	0.82	0.833 ± 0.05	0.846	0.822	0.834 ± 0.03
Relation classification	0.883	0.834	0.858 ± 0.04	0.884	0.831	0.857 ± 0.03	0.888	0.855	0.872 ± 0.05
Relation extraction	0.684	0.574	0.624 ± 0.03	0.673	0.635	0.653 ± 0.03	0.696	0.632	0.662 ± 0.02

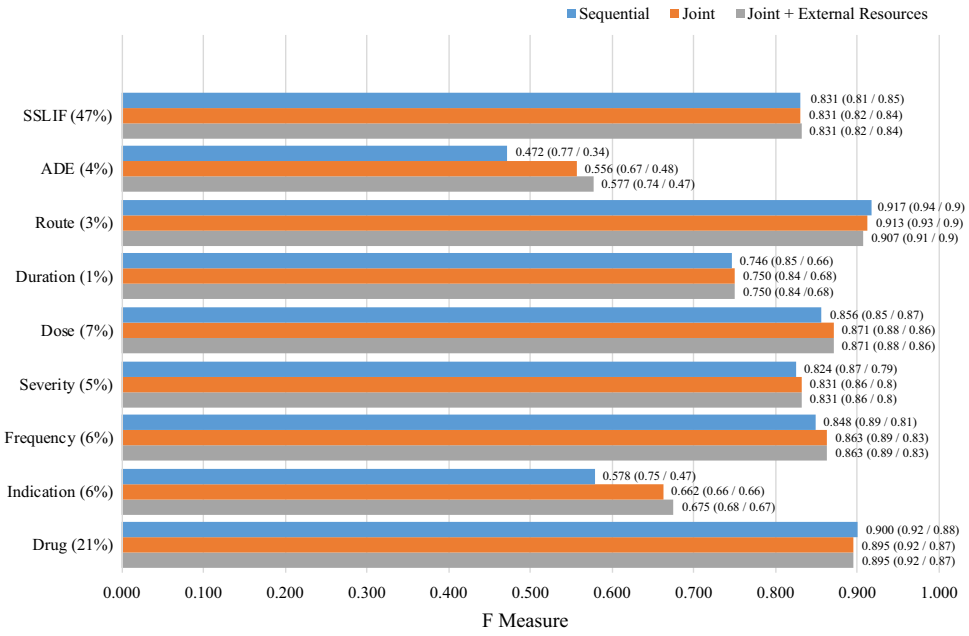

Fig. 6 Results for the entity extraction

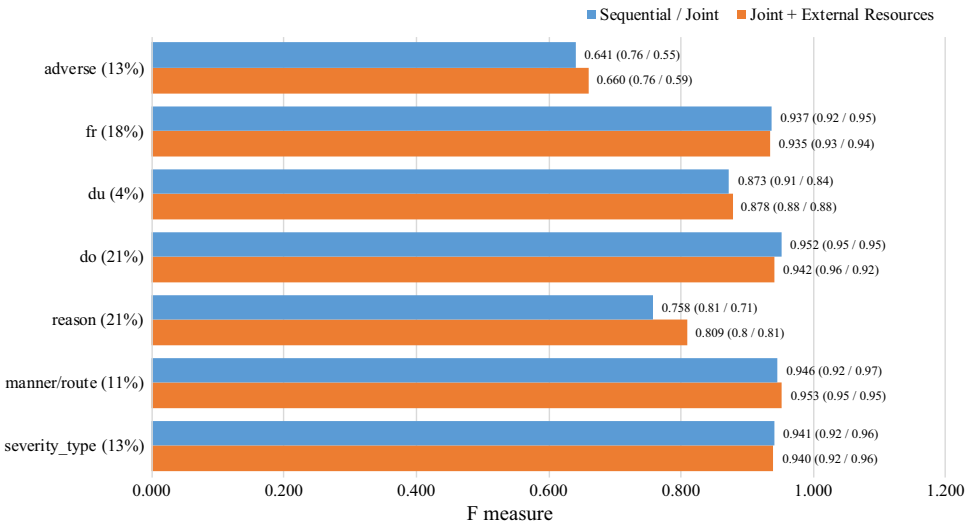

Fig. 7 Results for the relation extraction

was much lower than for the other classes. Poor performance of this method on the integrated task (see Fig. 8) is directly attributable to its low performance in recognizing ADEs and indications.

3.3 Joint Modeling

As introduced in Sect. 2.5, as a next step, we tried to improve upon the performance of the entity extraction by incorporating the existence of relations (or lack thereof) between entities. Overall, the micro-averaged *F* measure of this method was 0.833 for the entity extraction, 0.857 for the relation classification, and 0.653 for the integrated task. Performance improved by a relative 4.5% for the integrated task when compared with the sequential model. Entity recognition of ADEs and indications improved by a relative 13% and 14.5%, respectively.

3.4 Joint Modeling Plus External Resources

The best performance was achieved with the joint modeling plus external resources method, i.e., *F* measures of 0.834 for entity extraction, 0.872 for relation classification, and 0.662 for the integrated task, thus indicating the importance of incorporating domain knowledge for identifying adverse and reason relations, and in turn ADE and indication labels. Specifically, performance of ADE and indication extraction

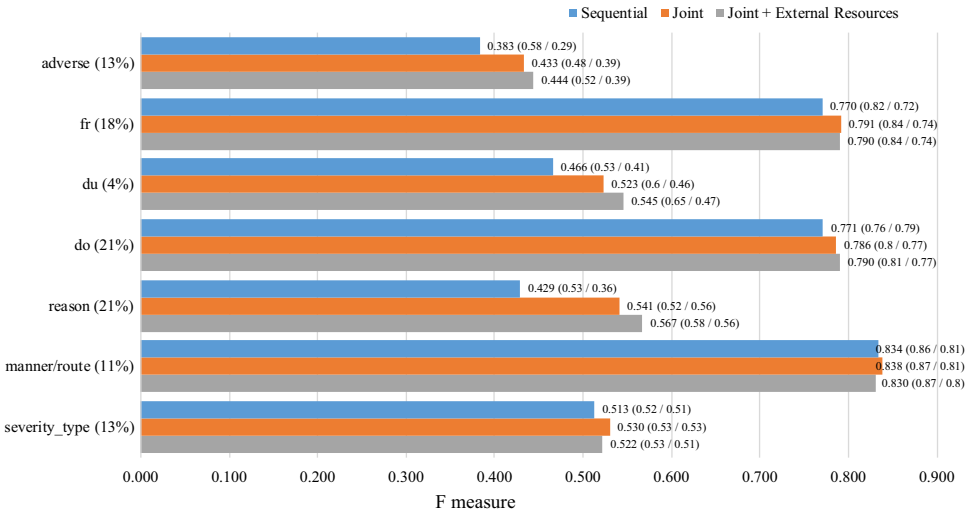

Fig. 8 Results for the integrated task

improved by a relative 4% and 2%, respectively, when compared with the joint model only.

3.5 Error Analysis

To gain further insights about our best-performing model, we conducted an error analysis (see Table 6). A major category of errors resulted from abbreviations and mis-spelled words, which are well-known in biomedical text processing. At least two other categories of errors resulted from the complexity of language processing. For example, "bleomycin toxicity" could be an SSLIF when considered as a single phrase or it could be two concepts—bleomycin (a drug) and toxicity (an SSLIF). The gold label annotation preferred the latter, whereas the system identified the former. The context appears to indicate that the system prediction was correct, but the system was penalized, nevertheless. Another frequent category of errors resulted from ambiguity in English words (e.g., emend has two meanings: it is the brand name of the drug aprepitant and also means to make corrections). Finally, we observed our system frequently misclassified the use of coordinating conjunctions (e.g., "left or right ventricular obstruction" was misclassified as "right ventricular obstruction").

4 Discussion

The importance of deep-learning-based approaches is evident in that most of the submissions in the MADE 1.0 challenge used variations of deep neural networks rather

Table 6 Error analysis

Error category	Examples	Gold labels	Predicted	Explanation
Abbreviations	Lymph node biopsy under **GETA** Bactrim 160 mg of **TMP** component **HPV** was negative	GETA—drug TMP—drug HPV—SSLIF		System misclassified rare/ambiguous abbreviations
Combination abbreviations	**OPEA** × 2 cycles **COPDAC** × 2 cycles	OPEA—drug COPDAC—drug		System misclassified combination abbreviations
Ambiguous terms	Continue with **Emend** for 2 days	Emend—drug		Emend as a word in English means "make corrections"
Spelling errors	**Vidodin** caused nausea Allergies: **prilose** statin	Vidodin—drug Prilose—drug		Spelling errors (should be Vicodin and Prilosec)
Phrase splitting	History of **bleomycin toxicity**	Bleomycin—drug Toxicity—SSLIF	Bleomycin toxicity—SSLIF	System predicted injury or poisoning caused by an external agent (SSLIF)
Coordinating conjunctions	No **left or right ventricular obstruction**	Left or right ventricular obstruction—SSLIF	Right ventricular obstruction—SSLIF	System misclassified long entities connected by a coordinating conjunction

GETA general anesthesia, *SSLIF* other signs or symptoms

Table 7 Performance comparison of our system and the next top two systems

Task	Chapman et al. [37]			Xu et al. [38]			Joint + external sources (our best system)		
	Precision	Recall	$F-1$	Precision	Recall	$F-1$	Precision	Recall	$F-1$
Concept extraction	0.838	0.781	0.809	0.842	0.827	0.816	0.846	0.822	0.834
Relation classification	*NA*	*NA*	0.868	*NA*	*NA*	0.832	0.888	0.855	0.872
Relation extraction	*NA*	*NA*	0.592	*NA*	*NA*	0.599	0.696	0.632	0.662

NA not available

than feature-based learning approaches. Several previous studies [2, 33, 34] demonstrated the need for LSTM-based networks for automated clinical entity recognition and relation extraction. Thus, as a first step, we implemented a baseline system that relied on LSTM-based networks, i.e., BiLSTM-CRF for entity recognition and attention-BiLSTM for relation classification. In this baseline system, we observed the importance of morphological, lexical and syntactical features as well as pre-trained embeddings. We made observations similar to previously reported results regarding the importance of using all three types of features as well as pre-trained embeddings for initializing the model inputs [33, 35, 36].

Most previous studies were on newswire articles and not on biomedical text, so results cannot be directly compared. A recent study by Li et al. [36] also used the joint modeling approach on biomedical text, but several critical differences exist between the studies. In Li et al. [36], manually summarized single sentences that were written in a textbook style were analyzed, meaning that the data analyzed were fundamentally different. The lexical scope of relations in the study was always within a sentence, whereas, here the scope of relations was an arbitrary number of sentences. Unlike Li et al. [36], we used the attention mechanism and knowledge-driven features, which improved the system's performance.

Several teams that participated in the MADE 1.0 challenge also analyzed clinical text. Table 7 compares the performance of our best method with the next two top-performing systems (as at the time of the challenge) for the integrated task. The systems used a method similar to our sequential approach but different machine learning models. Chapman et al. [37] employed a CRF model for concept extraction followed by a random forest model for relation extraction. Xu et al. [38] used BiLSTM-CRF for medical NER and support vector machine (SVM)-based pairwise relation classification between medical entities. Our method outperformed these two systems, indicating the effectiveness of (1) the state-of-the-art deep-learning models, (2) tailored methodologies to handle clinical text, (3) joint modeling of concept and relation extraction, and (4) knowledge-driven features.

5 Conclusions

We have reported our experience and results using state-of-the-art deep-learning neural networks for identifying entities and relations relevant to ADEs. We developed and assessed the performance of three methods using the neural networks: (1) a method that sequentially models entities first and then relevant relations among them; (2) a method that jointly models relations and certain key entities, leveraging the fact that the type of entities involved in a relation are predetermined; (3) a method where the information from external resources such as FAERS is used as an additional input to the neural networks. The methods provided increasing accuracy of the entity extraction and relation identification tasks, with the joint modeling plus external resources technique adding nearly 4 percentage points (or 6% relative improvement) to the current state of the art. The results from the second method were submitted to the MADE 1.0 challenge, where our system finished in first place in the overall integrated task and second in individual entity extraction and relation identification tasks.

Despite our success in the MADE challenge, there remains room for further improvement. Thus, in the future, we plan to explore several interesting research directions:

- *Joint inference for concept and relation extraction tasks* To overcome the error proportion problem in pipeline approaches for concept and relation extraction, we propose to explore joint inference models, which can make predictions for both tasks simultaneously.
- *Representation* Handling nested concepts (about 1% in this dataset), where span of one or more concepts overlaps with each other, with more advanced neural layered models for nested NER.
- *Incorporating external knowledge in concept and relation extraction* We plan to build embeddings from knowledge bases such as UMLS and use them in concept extraction and in the knowledge layer of relation extraction.
- *N-ary relation extraction using graph LSTMs* Explore a general framework for cross-sentence n-ary relation extraction based on graph LSTM networks.
- We plan to study the use of domain (EHR)-adapted dependency parsers to improve accuracy through better parsing of clinical text.

Compliance with Ethical Standards

Funding No sources of funding were used to conduct this study or prepare this manuscript.

Approval and consent This study was conducted on de-identified clinical notes as part of a shared challenge, so no ethical approval or patient consent was required.

Conflict of interest Bharath Dandala, Venkata Joopudi, and Murthy Devarakonda have no conflicts of interest that are directly relevant to the content of this article. Dr. Devarakonda is now on the faculty in Biomedical Informatics at Arizona State University, USA.

References

1. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. 2015. arXiv:1508.01991.
2. Chalapathy R, Borzeshi EZ, Piccardi M. Bidirectional LSTM-CRF for clinical concept extraction. In: Procedings of the clinical natural language processing workshop. 2016. pp. 7–12.
3. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017;33(14):i37–48.
4. Li F, Zhang M, Tian B, Chen B, Fu G, Ji D. Recognizing irregular entities in biomedical text via deep neural networks. Pattern Recognit Lett. 2017;105:105–13.
5. Dandala B, Mahajan D, Devarakonda M. IBM research system at TAC 2017: adverse drug reactions extraction from drug labels. In: Text analysis conference (TAC) 2017 workshop at NIST. 2017.
6. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2016. arXiv:1409.0473.
7. Hermann KM, et al. Teaching machines to read and comprehend. In: NIPS'15 proceedings of the 28th international conference on neural information processing systems, vol. 1. 2015. pp. 1693–1701.
8. Zhou P et al. Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers); 2016. pp. 207–212.
9. UMass BioNLP. NLP challenges for detecting medication and adverse drug events from electronic health records (MADE 1.0). https://bionlp.org/index.php/projects/39-nlp-challenges. Accessed 5 Feb 2018.
10. US Food and Drug Administration, "FAERS". https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm. Accessed 7 Feb 2018.
11. Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th annual meeting of the association for computational linguistics. 2016. pp. 1064–1074.
12. Zhang D, Wang D. Relation classification via recurrent neural network. 2015. arXiv:1508.01006.
13. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
14. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157–66.
15. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005;18(5):602–10.
16. Graves A. Generating sequences with recurrent neural networks. 2014. arXiv:1308.0850.
17. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT. 2016. pp. 260–270.
18. Collobert R, et al. Natural language processing (almost) from scratch. J Mach Learn Res. 2011;12:2493–537.
19. Sutton C, McCallum A, et al. An introduction to conditional random fields. Found Trends Mach Learn. 2012;4(4):267–373.
20. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in neural information processing systems, vol. 28. New York: Curran Associates Inc.; 2015. pp. 649–57.
21. Swampillai K, Stevenson M. Extracting relations within and across sentences. Proc Int Conf Recent Adv Nat Lang Process. 2011;2011:25–32.
22. Quirk C, Poon H. Distant supervision for relation extraction beyond the sentence boundary. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics. 2016. pp. 1171–1182.
23. Peng N, Poon H, Quirk C, Toutanova K, Yih W. Cross-sentence n-ary relation extraction with graph lstms. Trans Assoc Comput Linguis. 2017;5:101–15.
24. Griffis D, Shivade C, Fosler-Lussier E, Lai AM. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. AMIA Summits Transl Sci Proc. 2016;2016:88.
25. McCord MC, Bernth A. Using slot grammar. IBM TJ Watson Research Center, Yorktown Heights, NY, IBM Research Reports RC23978; 2010.
26. Minsky M. Memoir on inventing the confocal scanning microscope. Scanning. 1988;10(4):128–38.
27. Fillmore CJ. Frame semantics and the nature of language. Ann N Y Acad Sci. 1976;280(1):20–32.
28. Dandala B, Devarakonda M, Bornea M, Nielson C. Scoring disease-medication associations using advanced NLP, machine learning, and multiple content sources. In: Proceedings of the fifth workshop on building and evaluating resources for biomedical text mining (BioTxtM 2016). 2016. pp. 125–133.
29. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data. 2016;3:1–11.
30. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1):D267–70.
31. Rajani NF, Bornea M, Barker K. Stacking with auxiliary features for entity linking in the medical domain. BioNLP. 2017;2017:39–47.
32. Kingma D, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd international conference on learning representations (ICLR). 2015. pp. 1–15.
33. Jagannatha AN, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. In: Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing, vol. 2016. 2016. p. 856.
34. Munkhdalai T, Liu F, Yu H. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. JMIR Public Health Surveill. 2018;4(2):e29.
35. Sahu SK, Anand A, Oruganty K, Gattu M. Relation extraction from clinical texts using domain invariant convolutional neural network. In: Proceedings of the 15th workshop on biomedical natural language processing. 2016. pp. 206–215.
36. Li F, Zhang M, Fu G, Ji D. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinform. 2017;18(1):1–11.
37. Chapman AB, Peterson KS, Alba PR, DuVall SL, Patterson OV. Hybrid system for adverse drug event detection. In: Proceedings of machine learning research, vol. 90. 2018. pp. 16–24.
38. Xu D, Yadav V, Bethard S. UArizona at the MADE 1.0 NLP challenge. In: Proceedings of first international workshop on medication and adverse drug event detection. 2018. vol. 90, pp. 57–65.