Drug Safety (2019) 42:147–156
https://doi.org/10.1007/s40264-018-0763-y

ORIGINAL RESEARCH ARTICLE

Detecting Adverse Drug Events with Rapidly Trained Classification Models

Alec B. Chapman[1] · Kelly S. Peterson[2,3] · Patrick R. Alba[2,3] · Scott L. DuVall[2,3] · Olga V. Patterson[2,3]

Published online: 16 January 2019

Abstract

Introduction Identifying occurrences of medication side effects and adverse drug events (ADEs) is an important and challenging task because they are frequently only mentioned in clinical narrative and are not formally reported.
Methods We developed a natural language processing (NLP) system that aims to identify mentions of symptoms and drugs in clinical notes and label the relationship between the mentions as indications or ADEs. The system leverages an existing word embeddings model with induced word clusters for dimensionality reduction. It employs a conditional random field (CRF) model for named entity recognition (NER) and a random forest model for relation extraction (RE).
Results Final performance of each model was evaluated separately and then combined on a manually annotated evaluation set. The micro-averaged F1 score was 80.9% for NER, 88.1% for RE, and 61.2% for the integrated systems. Outputs from our systems were submitted to the NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0) competition (Yu et al. in http://bio-nlp.org/index.php/projects/39-nlp-challenges, 2018). System performance was evaluated in three tasks (NER, RE, and complete system) with multiple teams submitting output from their systems for each task. Our RE system placed first in Task 2 of the challenge and our integrated system achieved third place in Task 3.
Conclusion Adding to the growing number of publications that utilize NLP to detect occurrences of ADEs, our study illustrates the benefits of employing innovative feature engineering.

Key Points

Narrative clinical notes in electronic health records are frequently the only documentation of an occurred adverse drug event (ADE).

Natural language processing (NLP) can be employed to identify mentions of drugs and symptoms to facilitate detection of ADE mentions in clinical text.

While still an active area of research, progress is made in improving methods for NLP for ADE mention detection using advanced algorithms.

1 Introduction

Pharmacovigilance is a broad spectrum of activities that focus on identifying and preventing adverse drug events (ADEs), as well as understanding the risk factors and causes of ADEs when they do occur [2]. Relying on ADEs formally and

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.

✉ Olga V. Patterson
olga.patterson@utah.edu

[1] Health Fidelity, San Mateo, CA, USA

[2] VA Salt Lake City Health Care System, University of Utah, Salt Lake City, UT, USA

[3] Division of Epidemiology, University of Utah, Salt Lake City, UT, USA

spontaneously reported to the Food and Drug Administration (FDA) will inevitably lead to underestimating the risks imposed by medications [3]. In an effort to improve patient safety as well as mitigating risks, healthcare organizations have started implementing automated ADE detecting systems in electronic health records (EHRs) [4]. It has been long recognized that clinical narrative is the best source of information about suspected events related to medication [5]. While structured data in EHRs typically contain prescription and fill information for medications, as well as coded diagnoses, clinical narratives often provide descriptions of relationships between these concepts, such as a medicine prescribed to treat a condition or a side effect or ADE that may have occurred because of treatment. A wide variety of natural language processing (NLP) approaches have been previously explored in order to discover relationships between drugs and symptoms in EHRs as well as to learn about potential risks from biomedical literature [6, 7]. Despite progress in development of text processing techniques, clinical narrative continues to be an underutilized source of data for identifying unreported ADEs. Language variability as well as local environmental differences between different clinical settings limit adoption of NLP solutions across organizational boundaries [8].

Powerful machine learning algorithms based on deep learning, such as recurrent neural network (RNN) and convolutional neural network (CNN) that use pre-trained word embeddings [9, 10], have shown great results in their ability to capture complex relationships between concepts in text without effortful feature engineering [11]. RNN models have been accepted as the current state-of-the-art approach to labeling sequential data. While often high performing, training RNN is a computationally intensive process that takes time and possibly specialized hardware such as a graphic processing unit (GPU) [12]. Healthcare organizations as well as clinical research teams frequently lack the computational infrastructure needed for implementation of the latest text processing techniques, thus limiting their adoption [13]. Therefore, despite great advances in availability of high-performance computing infrastructures, it is essential to develop NLP systems that are fast, accurate, easily trainable in a new domain, and do not require specialized hardware.

We present a system that automatically identifies ADEs explicitly stated in clinical narratives as well as other information about patient drug treatments as submitted to the NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0) and our methods for all three tasks of the challenge [1, 14].

2 Methods

2.1 Data

A research team at the University of Massachusetts (UMASS) Medical School organized a shared task to tackle the problem of accurate detection of adverse drug events in clinical narrative. Detailed description of the shared task is presented elsewhere [14]. The shared task organizers prepared a set of de-identified clinical notes from the UMASS hospital and manually annotated them with the following categories:

- Drug, defined as any mention of a medication, including brand and generic names, as well as frequently used abbreviations.
- Indication, defined as a symptom that is a reason for drug administration.
- Adverse Drug Event (ADE), defined as a sign or symptom that resulted from a drug.
- Other Signs, Symptoms and Diseases (SSLIF), defined as any other sign or symptom that is not directly related to any drug mentioned in the note.
- Drug Frequency, defined as prescribed or suggested frequency of drug administration, such as ‘once per day’, or ‘as needed’.
- Drug Dose, or Dosage, defined as the amount of drug administered at one time.
- Drug Duration, defined as the length of time of a single prescription episode, such as ‘for 10 days’, or ‘for 2 weeks’.
- Drug Route, defined as the mode of administering the medication, such as ‘oral’, or ‘intravenous’.
- Severity, defined as the extent the disease or symptom affects the patient, such as ‘some’, or ‘severe’.

The annotated set also included relationships between different concepts (drugs, ADEs, indications, and signs and symptoms) that linked drug names to drug attributes (dose, route, frequency, and duration), drug names to ADEs, drug names to indications, and symptoms to severity.

The data set was split into two parts for training and testing of NLP systems and the sets were distributed to the participating teams at different times (see Table 1).

2.2 System Design

In accordance with a well established approach, our NLP system has two main modules: (1) identifying mentions of drugs, drug attributes, and symptoms as they are mentioned

Table 1 Concept instance distribution in training and testing sets

Concept	Instance count in training set	Instance count in testing set
Categories		
Drug	13,508	2395
Indication	3168	636
Frequency	4147	659
Severity	3374	534
Dose	4893	801
Duration	765	133
Route	2278	389
ADE	1509	431
SSLIF	34,056	5328
Relationships		
Severity	3476	559
Manner/route	2551	455
Reason	4554	876
Dosage	5177	866
Duration	906	147
Frequency	4419	730
Adverse	2082	530

ADE adverse drug event, *SSLIF* other signs, symptoms, and diseases

in clinical notes; and (2) classifying relationships between concepts with a set of predefined labels [6, 15].

2.2.1 Named Entity Recognition (NER)

Named Entity Recognition (NER) is a fundamental task in NLP that focuses on discovering mentions of a limited set of concept types. The traditional approach to the task is to use a predefined dictionary and expert-driven syntactic and semantic rules [16]. In the absence of a comprehensive dictionary for broad categories, statistical and supervised learning methods have been widely employed. Sequence-based classifier algorithms allow incorporating contextual information into the classification model. Deep learning algorithms for sequence-based classification are becoming increasingly popular for clinical NER because they alleviate the need for manual feature selection [17]. The main limitation of using neural networks is that model optimization typically involves hundreds or thousands of training iterations to perform hyperparameter search and cross-validation. While computational intensity of deep learning algorithms is widely recognized, challenges of working with such algorithms are frequently dismissed by citing availability of specialized hardware such as GPUs. In practice, while GPU acceleration aids in training neural network models, such hardware may not be available in all development and deployment environments. As Domingos writes, "machine learning is not a one shot process of building a dataset and running a learner, but rather an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating" [18]. When deciding on an appropriate algorithm, system designers have to consider the balance between the expert time and computational time. Due to the timing and resource constraints, we have selected to use a less computationally intensive algorithm and focus on feature engineering. Conditional random field (CRF) is a supervised machine learning classification algorithm that is simpler and, thus, faster training than RNN, but has the potential to perform well with minimal feature engineering [19]. The purpose of the constructed NER module was to label each token identified in clinical documents as one of the categories listed in Sect. 2.1.

Feature engineering is often a major step in building machine learning applications. One approach to representing words in text as numeric vectors is called word embeddings. The main benefit of word embeddings is that a model can be created from a large set of unlabeled documents and then reused for a variety of use cases [20]. For our system we used two sets of pretrained word embeddings as the basis for features. One set was trained as a continuous bag of words from public sources and nearly 100,000 EHR notes [21, 22]. Another set was trained as skip gram without any EHR data [23]. These sets are referred to as EHR and NoEHR embeddings in our system design description.

Following a previously described approach [20, 24–26], we included word embedding as cluster features rather than continuous values. The word embeddings vocabulary contained over 5 million features, therefore, we trained clusters with Mini-batch K-Means to work within available memory [27]. In addition, we included multiple cluster sizes ($K = 500$, 5000, 10000) and compound cluster features formed from token bigrams (e.g., "Cluster17_Cluster22") to capture generalizable phrases as opposed to strict bigrams as suggested by Guo et al. [20].

To identify known medications, the system included a lexicon of drug names using resources from MedEx [28]. This lexicon was then used to perform term matching in both local windows and in the entire sentence context.

The NER processing contained the following steps:

1. Sentence splitting using both a limited set of custom regular expressions and the Natural Language Toolkit (NLTK) [29].
2. Tokenization and part of speech (POS) labeling using NLTK.
3. Detecting mentions of known drug names using a lexicon developed from MedEx resources.
4. Building feature vectors from a variety of features.

The set of features included in the final model are as follows:

- Local features (window = 2 tokens):
  - Token, stem, POS tag
  - Patterns of capitalization, digits, and punctuation
  - Prefix and suffix characters (*n* = 2, 3)
  - Embedding clusters from unigrams and bigrams
  - Drug lexicon match
- Sentence features:
  - Drug lexicon match to the left or right of the current word

We utilized an annotated set of 876 clinical notes provided by MADE 1.0 organizers for training a CRF model for the NER module of the ADE detection system. The CRF model was trained using CRFSuite via the sklearn-crfsuite package available for scikit-learn [30, 31].

2.2.2 Relation Extraction (RE)

Once entities are detected in clinical documents, appropriate entities have to be linked in a relationship that represents the connection between these entities. For our system, relationships had to be identified between drug names and drug attributes—duration, route, frequency, and dosage; between drug names and symptoms that they caused—ADEs (labeled as Adverse), or that are reasons for prescription—indications (labeled as Reason); and between symptoms and severity concepts (labeled as Severity). Building the RE module was treated as a traditional supervised classification problem. We utilized features suggested by [32, 33]. Specifically, we extracted three types of features:

- Candidate Entities: Information about pairs of entities being considered for a relation:
  - Entity types
  - Entity word forms
- Entities Between: Other entities that appear between candidates
  - Entity types
  - Number of entities
- Surface Features: Tokens and POS tags between and neighboring the candidate entities
  - *N*-grams (*n* = 1–3)
  - Window size (1–3)
  - Number of tokens.

We divided RE into two subtasks: first, relation detection, which is a binary classification of whether any sort of relation exists between two entities; and second, relation classification, in which we classify what specific relation type exists [34]. Using a binary model for the first subtask helps to remove a number of false relations and improves classification precision. The multi-class classifier used by the second subtask is applied to all candidate pairs that were predicted to have a relation. Both classifiers are random forest models implemented in scikit-learn [31].

2.2.3 Full System

The integrated system combined NER and RE into a single pipeline with no additional processing. Source text is processed by the NER system preparing documents in BioC format [35], which the RE system augments with predicted relations.

3 Results

The NLP system validation was performed against the evaluation set provided by the MADE 1.0 challenge and our system performance was compared with performances of other submitted systems. The MADE 1.0 challenge defined a distinction between ‘standard’ and ‘extended’ resources employed by designed systems. Standard resources included only data resources provided by the challenge organizers, which were the EHR trained word embeddings. Any other resources could be used in the system design as extended. The final challenge results were initially reported on standard resources only; however, we also share findings when additional resources were used (e.g., NoEHR embeddings and MedEx) to illustrate how these resources improved performance.

The challenge was organized as three tasks: (1) NER, (2) RE, and (3) full system. The evaluation set contained 213 annotated documents that were used to obtain the validation results. Final performance of each model was evaluated separately and then combined on the evaluation. The micro-averaged F1 score was 80.9% for NER, 88.1% for RE, and 61.2% for the final system. During development and system training, a hold-out set containing 20% of the training data was used to evaluate the feature contribution for the RE model as well as detailed error analysis for the two main modules. The full evaluation set was used to evaluate the NER model feature contribution and error analysis.

3.1 Named Entity Recognition (NER) Results

Overall and per-label performance for our optimal NER model is presented in Table 2 while Table 3 summarizes the contributions from each feature class in the NER model.

Table 2 Performance metrics of the CRF NER model on the 213 final evaluation documents

Label	Standard resources			Extended resources		
	Precision	Recall	F1	Precision	Recall	F1
Route	94.4	87.4	90.8	94.8	89.5	92.1
Drug	90.1	83.7	86.8	91.1	86.1	88.6
Dose	89.2	83.5	86.3	89.8	85.4	87.5
Frequency	85.2	79.2	82.1	88.7	83.2	85.8
Severity	86.8	77.3	81.8	87.3	75.7	81.0
SSLIF	79.1	80.2	79.7	80.1	80.4	80.2
Duration	75.4	69.1	72.2	74.6	68.4	71.4
ADE	75.7	32.5	45.5	75.8	38.5	51.1
Indication	63.2	33.8	44.1	67.0	38.7	49.1
Overall micro	82.8	76.7	79.6	83.8	78.1	80.9

ADE adverse drug event, *CRF* conditional random field, *NER* named entity recognition, *SSLIF* other signs, symptoms, and diagnoses

Table 3 Contribution of NER model features by strict (exact text match) micro-averaged metrics

Features	Precision	Recall	F1
Baseline	82.1	71.4	76.4
+ Character features	75.6	74.6	77.9
+ Drug features	83.1	74.0	78.3
+ EHR embedding clusters (extended)	82.6	75.2	78.7
+ NoEHR embedding clusters (extended)	82.1	75.6	78.7
+ EHR and NoEHR embedding clusters (extended)	82.6	76.4	79.3
+ All features (standard)	82.8	76.7	79.6
+ All features (extended)	83.8	78.1	80.9

Baseline features were comprised of commonly used NER features such as tokens, stems, parts of speech and lexical patterns of capitalization, digits, and punctuation

EHR electronic health record, *NER* named entity recognition

Table 4 Comparison of training time between our system and the top performing submission in the NER task

Model type	Team	Time per training iteration (CPU)
Bi-LSTM-CRF	Worcester Polytechnic Institute [37]	480–3600 min
CRF	University of Utah	23 min

Bi bidirectional, *CRF* conditional random field, *LSTM* long short-term memory, *NER* named entity recognition

Performance was lowest on the ADE and Indication labels where recall was much lower than in the other classes.

Besides optimizing for F1, one of our objectives in using a CRF model was to allow rapid development of features and reduced training times. Wall time on CPU for extracting features for over 800 documents was measured at 2.5 min. In the system submitted to the MADE challenge, the optimizer for the training algorithm was L-BFGS [36].

Following the challenge, we corresponded with some of the top performing teams on the NER task. Since many of them used some form of RNN, we wanted to compare the time required to train our respective models. One example of training time comparison between our CRF model and the top performing system is shown in Table 4.

While the top performing team reported that one training fold of their model required approximately 4 h on GPU, we make a rough comparison by estimating the range of CPU training time from reported figures of 2 × to 15 × increase in training time [12, 38].

3.2 RE Results

Per-label performance using the final 213 evaluation documents for the RE model is shown in Table 5. Performance was lowest on 'Adverse' and 'Reason'. We performed additional analysis using an initial hold-out set of 176 documents from the training set. The contribution of each feature set is shown in Table 6.

3.3 Integrated System Results

The results for the final system are shown in Table 7.

Table 5 Performance metrics of the relation extraction model on the final 213 evaluation documents

Relation category	Precision	Recall	F1
Dosage	95.7	96.2	96.0
Frequency	97.1	92.3	94.7
Route	96.1	92.1	94.1
Severity	91.1	96.2	93.6
Duration	93.7	91.2	92.4
Reason	78.0	73.9	75.8
Adverse	78.7	68.3	73.1
Overall micro	90.3	85.9	88.1

Table 6 Contribution of features for the relation extraction model using a hold-out set of 176 documents

Features	Precision	Recall	F1
Entities between candidates	28.4	35.4	31.5
Candidate entities	42.7	72.8	53.9
Surface	74.6	66.2	70.2
Candidate entities + other entities between	81.6	90.4	85.8
All features	91.7	91.2	91.4

Table 7 Micro-averaged performance metrics of the final integrated model on the final 213 evaluation documents

Relation category	Precision	Recall	F1
Overall micro	72.1	53.4	61.2

3.4 Error Analysis

As the overall micro-averaged F1 score of the NER is relatively similar to the performance of other submissions, an error analysis was performed on the false negatives and positives on the ADE and Indication labels to categorize its incorrect predictions. We have identified the categories of errors starting with the most common in Table 8. Table 9 outlines the main categories of errors found when evaluating accuracy of relationship classification.

4 Discussion

Table 10 shows our final F1 scores on the 213 evaluation documents as reported by MADE 1.0 organizers using standard resources only. Table 11 shows the scores of our original test submissions alongside the three highest-performing submissions. Our system was ranked first in Task 2 and third in Task 3. Our scores have improved in each task since test submission. The score for the NER system was improved by incorporating extended resources that were not considered in reporting of top submissions. The score for the RE system was improved by fixing an error with sampling techniques during training.

A useful contribution of our approach to building an NER model is that it can be trained relatively quickly on commonly available hardware compared with neural network approaches. Noting that training times in Table 4 reflect a single fold of training, model optimization is clearly a highly computationally intensive task. Completing model optimization requires either long running processes on a single compute node or resources such as compute clusters or cloud

Table 8 Error analysis from NER predictions related to ADE and indication labels

Error category	Example	Explanation
Mislabeled Indication when Drug is not mentioned	"Treating currently as if she had **lymphoma**"	Without a mention of a Drug, Indication was predicted as SSLIF
Mislabeled SSLIF when unrelated Drug is mentioned	"History of **lymphoma** and was previously admitted for unrelated transplant and received *aspirin* at that time"	SSLIF was predicted as Indication due to Drug used in other treatment
Mislabeled SSLIF when Drug is not mentioned	"DISCHARGE DIAGNOSIS: **Lymphoma**"	Unexplained error when SSLIF was labeled as Indication when there was no mention of a Drug or treatment
Misclassification in short sentences	"No *urinary symptoms*"	Sentence contains too few words and urinary symptoms was incorrectly predicted as SSLIF
New note formatting	"**ALLERGIES**: Patient reported no itching or symptoms with the medication"	Allergy section format is different from training data, and ADE label was not assigned
Inconsistent prediction in a list	"Discussed potential side effects which include headaches, nausea, *vomiting*, diarrhea"	Unexplained error when vomiting was predicted as SSLIF while the others were correctly predicted as ADE
Contraindication mislabeledf as ADE	"Do not want to put her back on **colchicine** because of her *peripheral neuropathy*"	Contraindication diagnosis was predicted as ADE when Drug is mentioned

ADE adverse drug event, *NER* named entity recognition, *SSLIF* other signs, symptoms, and diseases

Table 9 Error analysis on relation extraction errors from a hold-out set of 176 documents

Error category	Example	Explanation
Implicit relation	"He did not have a fever with either cycles of ***chemotherapy***, but he did have 1 episode of *shingles*"	Drug was not explicitly stated to cause ADE
Entities more than two sentences away from each other	"50yo male with a *lymphoma*. …PLAN: 1…, 2. **Thalidomide** 50 mg a day"	Drug occurred in a different note section than Indication
Identical entity between first and second entity	"Her hematologist looking to initiate **erythropoietin**. I have discussed side effects of **erythropoietin** and would start weekly *injections*"	Another mention of identical Drug occurred closer to Route
Relation belongs to similar entity	"Patient received **lidocaine** and **hydrocortisone** *injection*"	A different Drug has Route
Historical treatment	"Patient presents for seventh cycle of hyper-CVAD for *mantle cell lymphoma*. Prior treatment consisted of **cyclophosphamide**"	Drug is not currently used as treatment for Indication
Annotation error	"Gabapentin 300 mg ***3 times daily***"	Frequency was not annotated with Drug

ADE adverse drug event

Table 10 Final evaluation scores for each task

	F1
Task 1—NER	79.6
Task 2—RE	88.1
Task 3—Integrated system	61.2

NER named entity recognition, *RE* relation extraction

computing services. Our findings suggest that employing deep learning techniques can be prohibitively expensive for a smaller research team with a short deadline. Using rapid feature engineering and training, we were able to quickly evaluate if our development efforts were successful.

Feature contribution shows that the NER model benefited from feature engineering including usage of a drug lexicon. Additionally, embedding cluster features improved performance where the optimal performance was achieved by employing both sets of pretrained embeddings, even though one embedding set did not include EHR documents in its training corpus.

The RE system performed best on categories such as 'Route', 'Frequency', and 'Dose', which are relatively simple statements that connect two entities that are often in close proximity in the text. The more challenging categories such as 'Reason' and 'Adverse' are often more linguistically complex and may involve some inference to understand that the two involved entities are connected. These categories will benefit from a more thorough analysis.

Feature engineering was an important component of the RE system. Of the three base feature sets that we considered, the surface features were by far the highest performing on the hold-out validation set. Although using only information about the entities being considered had a fairly low performance, adding information about what kinds of entities

Table 11 F1 scores reported by the MADE 1.0 organizers of the original test submissions

	Team name	References	Submission F1
Task 1—NER	Worcester Polytechnic Institute	[37]	82.9
	IBM Research	[39]	82.9
	University of Florida	[40]	82.3
	University of Utah		*79.6*
Task 2—RE	*University of Utah*		*86.8*
	IBM Research	[39]	84.0
	University of Arizona	[41]	83.2
Task 3—Integrated system	IBM Research	[39]	61.7
	University of Arizona	[41]	59.9
	University of Utah		*59.2*

The top three scores plus our score are shown for each task. Our scores are shown in *italics*

NER named entity recognition, *RE* relation extraction occur between them boosted performance considerably and resulted in a fairly competitive score. Using the union of all three resulted in the highest score.

The final integrated system combined both the NER and RE systems. The performance was significantly lower than the RE system using annotated documents, which shows the challenge of the integrated task.

Despite competing against more powerful and more computationally intensive approaches implemented by other submitted systems, our system achieved comparable results. The RE model was the top performing model in its task and the final system placed among the top three submissions.

4.1 Limitations

One limitation of the NER system is that the model assigns labels to SSLIF, ADE, and Indication in a single phase. Since these labels are all signs and symptoms with differing causality with respect to drugs, one possible improvement would be to establish a two-stage process. The first stage would combine the labels SSLIF/ADE/Indication into one label so that a second stage could disambiguate between these. Since the context window of the current CRF implementation is limited, the second stage could be a separate classifier that would use much more context than a single sentence to determine which label is the most appropriate. Features for this phase could include the current set but the features used in the RE system could also provide benefit. It would be interesting to see how such a staged architecture would perform compared with RNN models, which were the top performers in the challenge. Since feature engineering remained minimal, the CRF model would likely benefit from additional feature engineering for ADEs related in previous work [32]. One final limitation of the NER system that was seen during error analysis was that the current sentence detection algorithm was imperfect and often divided documents into sentences that were too small. Improved sentence breaking might particularly ameliorate the performance of ADE and Indication labels, as the current implementation limited the context available to one sentence, which was often far too short in size.

One other limitation of the integrated system was that we did not adjust either of the components (NER or RE) when combining them. Future work could focus on additional processing to improve the results when using both systems together.

Finally, since this challenge was conducted on a set of notes from oncology patients, it is unclear how well these models might generalize for pharmacovigilance in other medical domains. In future work, we intend to evaluate these models in the Department of Veterans Affairs to determine how well this work may translate to improving outcomes.

5 Conclusion

Automatic detection of adverse drug events can potentially have a profound effect on patient safety and accurate drug risk assessment. We developed a natural language processing system that can be retrained and applied in a new clinical setting without the use of specialized hardware, while still achieving performances comparable to more computationally intensive algorithms without requiring extensive feature engineering. Future work will include additional features and testing the system on a new dataset and in a new environment.

Source code for the NER system including feature extraction methods available at https://github.com/burgersmoke/MADE-CRF.

Acknowledgements This work was supported using resources and facilities at the VA Salt Lake City Health Care System with funding from VA Informatics and Computing Infrastructure (VINCI), VA HSR RES 13-457.

Compliance with Ethical Standards

Funding The work described in the manuscript was supported using resources and facilities at the VA Salt Lake City Health Care System. The work in part was funded by VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204.

Conflict of Interest Olga V. Patterson reports grants from National Heart, Lung, and Blood Institute, Department of Defense, Amgen Inc., Anolinx LLC, Genentech Inc., Gilead Sciences Inc., Merck & Co., Inc., Novartis International AG, and PAREXEL International Corporation outside the submitted work. Scott L. DuVall reports grants from National Heart, Lung, and Blood Institute during the conduct of the study; and grants from AbbVie Inc., Amgen Inc., Anolinx LLC, Astellas Pharma Inc., AstraZeneca Pharmaceuticals LP, Boehringer Ingelheim International GmbH, F. Hoffman-La Roche Ltd, Genentech Inc., Genomic Health, Inc., Gilead Sciences Inc., GlaxoSmithKline PLC, HITEKS Solutions Inc., Innocrin Pharmaceuticals Inc., Kantar Health, LexisNexis Risk Solutions, Merck & Co., Inc., Mylan Specialty LP, Myriad Genetics, Inc., Northrop Grumman Information Systems, Novartis International AG, PAREXEL International Corporation, and Shire PLC outside the submitted work. Alec B. Chapman, Kelly S. Peterson and Patrick R. Alba report no conflicts of interest that are directly relevant to the content of the reported study.

Ethical Approval The study has been approved by the University of Utah Institutional Review Board.

References

1. Yu H, Jagannatha AN, Liu F, Liu W. NLP Challenges for detecting medication and adverse drug events from electronic health records. 2018. http://bio-nlp.org/index.php/projects/39-nlp-challenges. Accessed 5 Dec 2018.
2. World Health Organization. WHO|Pharmacovigilance [Internet]. WHO. World Health Organization; 2015. http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/. Accessed 11 Sept 2018.
3. Alatawi YM, Hansen RA. Empirical estimation of under-reporting in the U.S. Food and Drug Administration Adverse Event Reporting System (FAERS). Expert Opin Drug Saf [Internet]. 2017;16(7):761–7. http://www.ncbi.nlm.nih.gov/pubmed/28447485. Accessed 30 Jun 2018.
4. Vilar S, Harpaz R, Santana L, Uriarte E, Friedman C. Enhancing adverse drug event detection in electronic health records using molecular structure similarity: application to pancreatitis. PLoS One. 2012;7(7):e41471. http://www.ncbi.nlm.nih.gov/pubmed/22911794.
5. Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inf Assoc. 2010;17(6):671–4. http://www.ncbi.nlm.nih.gov/pubmed/20962129.
6. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf. 2017;40(11):1075–89. http://www.ncbi.nlm.nih.gov/pubmed/28643174. Accessed 15 Apr 2018.
7. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text mining for adverse drug events: the promise, challenges, and state of the art. Drug Saf. 2014;37(10):777–90. http://www.ncbi.nlm.nih.gov/pubmed/25151493.
8. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc. 2017;24(5):986–91. https://doi.org/10.1093/jamia/ocx039.
9. Mikolov T, Sutskever I, Chen K, et al. GC-A in neural, 2013 U. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems [Internet]. Lake Tahoe, Nevada: Curran Associates Inc.; 2013. pp. 3111–9. http://papers.nips.cc/paper/5021-distributed-representations-of-words-andphrases. Accessed 13 Apr 2018.
10. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proc int conf learn represent (ICLR 2013) [Internet]. 2013;1–12. https://arxiv.org/abs/1301.3781. Accessed 15 Apr 2018.
11. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J Biomed Inform. 2017;76:102–9. http://arxiv.org/abs/1706.09569.
12. Li B, Zhou E, Huang B, Duan J, Wang Y, Xu N, et al. Large scale recurrent neural network on GPU. In: 2014 International joint conference on neural networks (IJCNN) [Internet]. IEEE;2014;4062–9. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=889433. Accessed 23 Apr 2018.
13. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):52. https://doi.org/10.1098/rsif.2017.0387.
14. Jagannatha A, Liu F, Liu W, Yu H. Overview of the first natural language processing challenge for extracting medication, indication and adverse drug events from electronic health record notes (MADE1.0). Drug Saf. 2018.
15. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform [Internet]. 2010;160(Pt 1):739–43. http://www.ncbi.nlm.nih.gov/pubmed/20841784.
16. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investig [Internet]. 2007;30(1):3–26. https://benjamins.com/catalog/li.30.1.03nad. Accessed 16 Sept 2018.
17. Wu Y, Jiang M, Xu J, Zhi D, Xu H. Clinical Named Entity Recognition Using Deep Learning Models. In: AMIA. Annu Symp proceedings AMIA Symp [Internet]. 2017:1812–9. http://www.ncbi.nlm.nih.gov/pubmed/29854252. Accessed 16 Sept 2018.
18. Domingos P. A few useful things to know about machine learning. Commun ACM [Internet]. 2012;55(10):78. http://dl.acm.org/citation.cfm?doid=2347736.2347755. Accessed 16 Sept 2018.
19. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML '01 Proc Eighteenth Int Conf Mach Learn [Internet]. 2001;8:282–9. https://repository.upenn.edu/cis_papers/159/. Accessed 06 Apr 2018.
20. Guo J, Che W, Wang H, Liu T. Revisiting Embedding Features for Simple Semi-supervised Learning. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) [Internet]. Association for computational linguistics; 2014. p. 110–20. http://www.aclweb.org/anthology/D/D14/D14-1012.pdf. Accessed 07 Mar 2018.
21. Jagannatha A, Yu H. Structured prediction models for RNN based sequence labeling in clinical text. In: The 2016 conference on empirical methods in natural language processing [Internet]. Austin, Texas: Association for computational linguistics; 2016. p. 856–65. https://www.aclweb.org/anthology/D16-1082. Accessed 07 Mar 2018.
22. Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. In: Proc Conf Assoc Comput Linguist North Am Chapter Meet [Internet]. 2016;2016:473–82. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5119627/. Accessed 06 Apr 2018.
23. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. In: Proc LBM [Internet]. 2013;39–44. https://pdfs.semanticscholar.org/e2f2/8568031e1902d4f8ee818261f0f2c20de6dd.pdf. Accessed 06 Apr 2018.
24. Turian J, Ratinov L, Meeting YB-P of the 48th annual, 2010 U. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics [Internet]. 2010. https://dl.acm.org/citation.cfm?id=1858721. Accessed 06 Apr 2018.
25. Yu M, Zhao T, Dong D, Tian H, Yu D. Compound embedding features for semi-supervised learning. In: Proc NAACL-HLT [Internet]. 2013;(June):563–8. http://www.aclweb.org/anthology/N13-1063. Accessed 06 Apr 2018.
26. Nikfarjam A, Sarker A, O'connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inf Assoc. 2015;22(3):671–81.
27. Sculley D. Web-scale k-means clustering. In: Proceedings of the 19th international conference on World wide web—WWW '10 [Internet]. New York, New York, USA: ACM Press; 2010. p. 1177. http://portal.acm.org/citation.cfm?doid=1772690.1772862. Accessed 23 Apr 2018.

28. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc [Internet]. 2010;17(1):19–24. http://jamia.bmj.com/content/17/1/19.abstract. Accessed 11 Aug 2011.
29. Bird S, Klein E, Loper E. In: Steele J, editor. Natural language processing with python. 1st ed. O'Reilly Media Inc.; 2009.
30. Okazaki N. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite. Accessed 5 Dec 2018.
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res [Internet]. 2012;12:2825–30. http://www.jmlr.org/papers/v12/pedregosa11a.html. Accessed 06 Apr 2018.
32. Liu J, Zhao S, Wang G. SSEL-ADE: a semi-supervised ensemble learning framework for extracting adverse drug events from social media. Artif Intell Med [Internet]. 2018;84:34–49. https://www.sciencedirect.com/science/article/pii/S0933365717301847. Accessed 06 Apr 2018.
33. GuoDong Z, Jian S, Jie Z, Min Z. Exploring various knowledge in relation extraction. In: Proceedings of the 43rd annual meeting on association for computational linguistics—ACL '05 [Internet]. Morristown, NJ, USA: association for computational linguistics; 2005. p. 427–34. http://portal.acm.org/citation.cfm?doid=1219840.1219893. Accessed 23 Apr 2018.
34. Kumar S. A survey of deep learning methods for relation extraction. arXiv Prepr arXiv170503645 [Internet]. 2017 May 10. http://arxiv.org/abs/1705.03645. Accessed 23 Apr 2018.
35. Comeau DC, Doğan RI, Ciccarese P, Cohen KB, Krallinger M, Leitner F, et al. BioC: a minimalist approach to interoperability for biomedical text processing. Database [Internet]. 2013;2013(0):bat064. http://www.ncbi.nlm.nih.gov/pubmed/24048470. Accessed 23 Apr 2018.
36. Nocedal J. Updating Quasi-Newton Matrices with Limited Storage. Math Comput [Internet]. 1980;35(151):773. http://www.ams.org/jourcgi/jour-getitem?pii=S0025-5718-1980-0572855-7. Accessed 01 May 2018.
37. Wunnava S, Qin X, Kakar T, Rundensteiner EA, Kong X. Bidirectional LSTM-CRF for adverse drug event tagging in electronic health records. Proc Mach Learn Res. 2018;90:48–56.
38. Raina R, Madhavan A, Ng AY. Large-scale deep unsupervised learning using graphics processors. In: Proceedings of the 26th annual international conference on machine learning. ACM; 2009. p. 873–80.
39. Dandala B, Joopudi V, Devarakonda M. IBM Research System at MADE 2018: Detecting adverse drug events from electronic health records. In: International Workshop on Medication and Adverse Drug Event Detection. 2018. p. 39–47.
40. Yang X, Bian J, Wu Y. Detecting medications and adverse drug events in clinical notes using recurrent neural networks. In: International workshop on medication and adverse drug event detection. 2018. p. 1–6.
41. Xu D, Yadav V, Bethard S. UArizona at the MADE 1.0 NLP Challenge. In: International workshop on medication and adverse drug event detection. 2018. p. 57–65.