Proceedings of Machine Learning Research 90:16–24, 2018 Medication and Adverse Drug Event Detection Workshop

Hybrid system for adverse drug event detection

Alec B. Chapman
Kelly S. Peterson
Patrick R. Alba
Scott L. DuVall
Olga V. Patterson OLGA.PATTERSON@UTAH.EDU
VA Salt Lake City Health Care System;
University of Utah, Salt Lake City, UT, USA

Editor: Feifan Liu, Abhyuday Jagannatha, Hong Yu

Abstract

In context of the NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records (MADE 1.0) (Yu et al., 2018), we built a hybrid natural language processing system that combined multiple algorithms and resources to identify the relationship between mentions of symptoms and drugs. Our system employed a conditional random field (CRF) model for named entity recognition (NER) and a random forest model for relation extraction (RE). Final performance of each model was evaluated separately and then combined on the challenge's hold-out evaluation set. The micro-averaged F1 score was 80.9% for NER, 86.8% for RE, and 59.2% for the final system.

Keywords: Natural Language Processing, Adverse Drug Events, Clinical Text.

1. Introduction

Pharmacovigilance is the practice of understanding patient risks associated with medical treatment and identifying and preventing adverse drug events (ADEs) when they do occur. One tool useful for identifying ADEs is the clinical narrative. In addition to prescription and fill information for medications and lists of diagnoses for conditions that may be stored as structured data in electronic health records (EHR), 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. Recognizing the type of the relationship between a medication and condition is an essential step in providing accurate data for health out-comes research, ADE reporting, and pharmacovigilance. 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 (Yu et al., 2018).

2. Methods

2.1. Named Entity Recognition (NER)

Recurrent Neural Network (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) (Li et al., 2014). We elected to experiment with conditional random field (CRF) models (Lafferty et al., 2001) to determine how well simpler, faster training models might perform with minimal feature engineering. The model was trained using CRFSuite via the sklearn-crfsuite package available for scikit-learn (Okazaki, 2007; Pedregosa et al., 2012). In accordance with previous findings (Turian et al., 2010; Yu et al., 2013; Guo et al., 2014), we included word embeddings as clusters rather than continuous values as features. The word embeddings vocabulary contained over 5 million features, therefore, we trained clusters with Mini-batch KMeans to handle such a large vocabulary (Sculley, 2010). 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. (2014).

Two sets of pretrained word2vec word embeddings were used for features (Mikolov et al., 2013a,b). One set was trained as continuous bag of words from public sources and nearly 100,000 EHR notes (Jagannatha and Yu, 2016a,b). Another set was trained as skip gram without any EHR data Pyysalo et al. (2013). These sets are referred to as EHR and NoEHR embeddings in our system design description.

Finally, a lexicon of drug names was implemented using resources from MedEx (Xu et al., 2010) to implement term matching in both local windows and entire sentence context. The set of features included in the final model are:

- Local features (window = 2):
  - Token, stem, part of speech
  - 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.

2.2. Relation Extraction (RE)

We treated the RE task as a traditional supervised classification problem. We utilize features suggested by Liu et al. (2018) and GuoDong et al. (2005). 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, where we classify what specific relation type exists (Kumar, 2017). The first task uses a binary model that classifies whether there is any sort of relation between two entities. This helps remove a number of false relations and improves classification precision. The second task uses a multi-class classifier that is applied to all candidate pairs that were predicted to have a relation. Both classifiers are random forest models implemented in scikit-learn (Pedregosa et al., 2012).

2.3. Full System

The full 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 (Comeau et al., 2013), which the RE system augments with predicted relations.

3. Results

The challenge was organized as three tasks: 1) NER, 2) RE, and 3) full system. For the NER task, the results are based on the test data that was used for evaluation. Because the reference standard has not been made available for Task 2, the results are obtained using a hold-out set containing 20% of the training data.

3.1. Named Entity Recognition (NER) results

Table 1 shows the contributions from each feature class in the NER model. Per-label performance for the optimal NER model is presented in Table 2. Performance was lowest on the ADE and Indication labels where recall was much lower than the other classes.

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 3.

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

Table 1: Contribution of NER model features by strict (exact text match) micro-averaged metrics. Baseline features were comprised of commonly used NER features such as tokens, stems, part-of-speech and lexical patterns of capitalization, digits and punctuation.

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	82.6	75.2	78.7
+ NoEHR Embedding Clusters	82.1	75.6	78.7
+ EHR and NoEHR Embedding Clusters	82.6	76.4	79.3
+ All features	83.8	78.1	80.9

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

Features	**Precision**	**Recall**	**F1**
Drug	91.1	86.1	88.6
Indication	67.0	38.7	49.1
Frequency	88.7	83.2	85.8
Severity	87.3	75.7	81.0
Dose	89.8	85.4	87.5
Duration	74.6	68.4	71.4
Route	94.8	89.5	92.1
ADE	75.8	38.5	51.1
SSLIF	80.1	80.4	80.2
Overall Micro	83.8	78.1	80.9

Table 3: 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 (Drug) 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 mislabeled as *ADE*	"Do not want to put her back on (Drug) because of her **peripheral neuropathy**"	Contraindication diagnosis was predicted as *ADE* when *Drug* is mentioned.

Table 4: Contribution of features for the RE 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 5: Performance metrics of the RE model using a hold-out set of 176 documents.

Features	**Precision**	**Recall**	**F1**
Severity Type	94.7	95.7	95.2
Manner/Route	97.2	97.4	97.3
Reason	79.6	79.4	79.5
Dosage	96.6	97.8	97.2
Duration	93.5	97.7	95.6
Frequency	96.3	97.6	96.9
Adverse	89.4	76.6	82.5
Overall Micro	91.7	91.2	91.4

features for over 800 documents was measured at 2.5 minutes and each training of 5-fold cross validation was 22.5 minutes. In all findings, the optimizer for the training algorithm was L-BFGS (Nocedal, 1980). It may be worth exploring other optimizers with regards to training time.

3.2. Relation Extraction (RE) Results

Results of features contributing in the RE model are shown in Table 4.

Per-label performance using a hold-out set of 176 documents for the RE model is shown in Table 5. Performance was lowest on "Adverse" and "Reason".

3.3. Full System

Final performance of each model was evaluated separately and then combined on the challenge's hold-out evaluation set. The micro-averaged F1 score was 80.9% for NER, 86.8% for RE, and 59.2% for the final system.

4. Discussion

A useful contribution of our NER model can be trained on commonly available hardware relatively quickly compared to neural network approaches. While GPU acceleration aids

Table 6: Error analysis on relation extraction errors.

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* occurs closer to *Manner/Route*.
Relation belongs to similar entity	"Patient received **lidocaine** and **hydrocortisone** injection."	A different *Drug* has *Manner/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*

neural network models, such hardware may not be available in all development and deployment environments. Feature contribution shows that the 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. The CRF model would likely benefit from additional feature engineering for ADEs related in previous work (Liu et al., 2018). Error analysis also showed opportunity to improve sentence breaking as the current implementation limited available context.

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 (F1=70.2). Although using only information about the entities being considered had a fairly low performance (F1=53.9), adding information about what kinds of entities occur between them boosted performance considerably and resulted in a fairly competitive score (F1=85.8). Using the union of all three resulted in the highest score (F1=91.4).

The RE system performed best on categories such as "Manner/Route", "Frequency", and "Dosage", 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.

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

Machine learning models offer effective tools for pharmacovigilance. Drug-related entities can be identified using sequence labeling methods and can then be linked together through utilizing machine learning classification methods.

Acknowledgments

This work was also 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.

References

Donald C. Comeau, Rezarta Islamaj Doan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. BioC: A minimalist approach to interoperability for biomedical text processing. *Database*, 2013(0):bat064–bat064, sep 2013.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Revisiting Embedding Features for Simple Semi-supervised Learning. In *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pages 110–120. Association for Computational Linguistics, 2014.

Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proc. 43rd Annu. Meet. Assoc. Comput. Linguist. - ACL '05*, pages 427–434, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

Abhyuday Jagannatha and Hong Yu. Structured prediction models for RNN based sequence labeling in clinical text. In *2016 Conf. Empir. Methods Nat. Lang. Process.*, pages 856–865, Austin, Texas, nov 2016a. Association for Computational Linguistics.

Abhyuday N Jagannatha and Hong Yu. Bidirectional RNN for Medical Event Detection in Electronic Health Records. *Proc. Conf. Assoc. Comput. Linguist. North Am. Chapter. Meet.*, pages 473–482, jun 2016b.

Shantanu Kumar. A Survey of Deep Learning Methods for Relation Extraction. *arXiv Prepr. arXiv1705.03645*, may 2017.

John Lafferty, Andrew McCallum, and Fernando C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01 Proc. Eighteenth Int. Conf. Mach. Learn.*, 8(June):282–289, 2001.

Boxun Li, Erjin Zhou, Bo Huang, Jiayi Duan, Yu Wang, Ningyi Xu, Jiaxing Zhang, and Huazhong Yang. Large scale recurrent neural network on GPU. In *2014 Int. Jt. Conf. Neural Networks*, pages 4062–4069. IEEE, jul 2014.

Jing Liu, Songzheng Zhao, and Gang Wang. SSEL-ADE: A semi-supervised ensemble learning framework for extracting adverse drug events from social media. *Artif. Intell. Med.*, 84:34–49, jan 2018.

T Mikolov, I Sutskever, K Chen, GS Corrado Advances in neural . . . , and Undefined 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *NIPS'13 Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, pages 3111–3119, Lake Tahoe, Nevada, 2013a. Curran Associates Inc.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proc. Int. Conf. Learn. Represent. (ICLR 2013)*, pages 1–12, jan 2013b.

Jorge Nocedal. Updating Quasi-Newton Matrices with Limited Storage. *Math. Comput.*, 35(151):773, sep 1980.

Naoaki Okazaki. CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007. URL http://www.chokkan.org/software/crfsuite/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, jan 2012.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. Distributional Semantics Resources for Biomedical Text Processing. *Proc. LBM*, pages 39–44, 2013.

D. Sculley. Web-scale k-means clustering. In *Proc. 19th Int. Conf. World wide web - WWW '10*, page 1177, New York, New York, USA, 2010. ACM Press.

J Turian, L Ratinov, Y Bengio Proceedings of the 48th annual Meeting, and Undefined 2010. Word representations: a simple and general method for semi-supervised learning. In *Proc. 48th Annu. Meet. Assoc. Comput. Linguist.*, 2010.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. MedEx: a medication information extraction system for clinical narratives. *J. Am. Med. Inform. Assoc.*, 17(1):19–24, 2010.

Hong Yu, Abhyuday N Jagannatha, Feifan Liu, and W Liu. NLP Challenges for Detecting Medication and Adverse Drug Events from Electronic Health Records, 2018. URL https://bio-nlp.org/index.php/announcements/39-nlp-challenges.

Mo Yu, Tiejun Zhao, Daxiang Dong, Hao Tian, and Dianhai Yu. Compound Embedding Features for Semi-supervised Learning. In *Proc. NAACL-HLT*, number June, pages 563–568, 2013.