Evaluation of Transfer Learning for Adverse Drug Event (ADE) and Medication Entity Extraction

Sankaran Narayanan  
Amrita Vishwa Vidyapeetham  
Amritapuri, India  
nsankaran@am.amrita.edu

Kaivalya Mannam  
Georgia Institute of Technology  
Atlanta, Georgia  
kmannam3@gatech.edu

Sreeranga P. Rajan  
Alphabet Inc. and Stanford University  
California, USA  
sree@cs.stanford.edu

P. Venkat Rangan  
Amrita Vishwa Vidyapeetham  
Amritapuri, India  
venkat@amrita.edu

Abstract

We evaluate several biomedical contextual embeddings (based on BERT, ELMo, and Flair) for the detection of medication entities such as Drugs and Adverse Drug Events (ADE) from Electronic Health Records (EHR) using the 2018 ADE and Medication Extraction (Track 2) n2c2 data-set. We identify best practices for transfer learning, such as language-model fine-tuning and scalar mix. Our transfer learning models achieve strong performance in the overall task (F1=92.91%) as well as in ADE identification (F1=53.08%). Flair-based embeddings out-perform in the identification of context-dependent entities such as ADE. BERT-based embeddings out-perform in recognizing clinical terminology such as Drug and Form entities. ELMo-based embeddings deliver competitive performance in all entities. We develop a sentence-augmentation method for enhanced ADE identification benefiting BERT-based and ELMo-based models by up to 3.13% in F1 gains. Finally, we show that a simple ensemble of these models outpaces most current methods in ADE extraction (F1=55.77%).

1 Introduction

Adverse Drug Events (ADE) arising from the medical intervention of drugs account for 1.3 million visits to the emergency department in the United States alone (CDC, 2017). Randomized controlled trials (RCTs), the primary mechanism for monitoring and identifying ADEs, are hampered by insufficient sample sizes of clinical trials (Sultana et al., 2013). Pharmacovigilance databases such as the Food and Drug Administration’s Adverse Event Reporting System (FAERS) strive to be authoritative sources for Physicians; however, they require regular manual data entry (Hoffman et al., 2014; Chedid et al., 2018).

Electronic Health Records (EHRs) contain valuable information about patient medication history: drugs prescribed, reasons for administration, dosages/strengths, and ADEs. Automated extraction of these medication entities by Natural Language Processing (NLP) techniques can facilitate wide-scale pharmacovigilance (Moore and Furberg, 2015; Liu et al., 2019a).

Incorporating such a predictive system within the clinical note-taking interface may help the Physician by alleviating the need to access external clinical decision support applications (Chen et al., 2016). For instance, if a physician notes down ‘started on Dilantin for seizure prophylaxis for a few days’, the text could be quickly parsed - highlighting ‘Dilantin’ as a drug, ‘seizure prophylaxis’ as the reason for administration, ‘few days’ as the duration, and warnings of ‘eye discharge’, ‘oral sores’, etc. as potential ADEs. In the example given, ‘seizure prophylaxis’ and ‘few days’ may occur any where in the clinical text, but only in the context of ‘Dilantin’ they indicate reason / duration for administration. Besides, such ‘dynamic’ interfaces can aid medical students to learn from their collective experiences.

Among medication entities, ADE and Reason are challenging to disambiguate (Henry et al., 2020). Frequently, the specific reason for drug administration may appear in a subsequent sentence (Dandala et al., 2020). Besides, ADE data-sets include gold-annotations for these entities, only if they are associated with a drug. Doing so leads to a significant reduction in the number of gold annotations (Wei et al., 2020).

As part of our work in uniting clinical decision support functions and note-taking interfaces, we needed to develop a high-performing medication extraction model using open-source NLP frameworks. Following (Miller et al., 2019), we modeled this as a named-entity recognition task (Uzuner...
| S.No. | Author                      | Method                                                                 | Overall F1 | ADE F1 |
|-------|-----------------------------|------------------------------------------------------------------------|------------|--------|
| 1.    | Alibaba Inc. (Henry et al., 2020) | BiLSTM-CRF + ELMo embedding, Section Features | 94.18      | 58.73  |
| 2.    | Dandala et al. (2020)       | BiLSTM-CRF Custom-trained ELMo using MIMIC-III Knowledge-embeddings from FAERS Custom pre-processing | 93.5       | 53.5   |
| 3.    | Wei et al. (2020)           | CRF + BiLSTM-CRF + Joint 3-model NER ensemble; joint-relation classifier | 93.45      | 52.95  |
| 4.    | Ju et al. (2020)            | 4-layer tree-structured BiLSTM-CRF Word, sub-word, and character embeddings Three-groups of specialized features Overlapping span handling | 92.55      | 27.90  |
| 5.    | Kim and Meystre (2020)      | CRF+CRFext+SEARN+BiLSTM ensemble Glove embeddings Inputs from MedEx and external corpora Stanford CoreNLP for tokenization | 92.66      | 27.11  |
| 6.    | Dai et al. (2020)           | CRF + BiLSTM-CRF Cascading BiLSTM architecture Pre-trained domain-specific embeddings Nested entity handling | 91.9       | 38.75  |
| 7.    | Miller et al. (2019)        | BiLSTM-CRF Flair embeddings (general purpose corpora) Default features and hyper-parameters *: 50 epoch run, final performance could be higher | 90*        | 27*    |
| 8.    | Chen et al. (2020)          | BiLSTM-CRF UMLS-based concept lookups Specialized handling of temporal entities Regular expressions and rules | 84.97      | 43.29  |

Table 1: Relevant related work.

et al., 2011; Si et al., 2019) and experimented with transfer learning using openly available biomedical contextual embeddings. It is in this context,

1. We evaluate transfer learning models incorporating: BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), ELMo (Peters et al., 2018) and Flair (Akbik et al., 2018) contextual embeddings pre-trained on PubMed abstracts (Fiorini et al., 2018).

2. We evaluate embedding-specific methods to maximize performance: language-model fine-tuning, scalar mix, sub-word token aggregation.

3. Based on the performance of the transfer learning models, we develop procedures for enhanced ADE and Reason identification. Sentence-augmentation at prediction-time benefits ADE extraction by up to +3.13% in F1 gains. It also facilitates a deeper understanding of the behavior of the embeddings. Ensembling strategies help improve performance of all three challenging entities: ADE, Duration, and Reason with up to +2.63% in F1 gains for ADE.

Our main intention was to get a transfer learning pipeline working with these embeddings and therefore we did not perform any detailed hyper-parameter optimization. Despite this, we were able to achieve strong performance with all the embeddings. Standalone models achieved F1-scores of 53.08% in ADE extraction and 92.91% in the overall task with default features. A basic ensemble constructed from these standalone models achieved F1-scores of 55.77% in ADE extraction and 92.82% in the overall task confirming the viability of the overall strategy.

2 Related Work

Classical research in this area focused on rule-based systems (such as MedEx (Xu et al., 2010), ADEPt (Iqbal et al., 2017)) and CRF-based machine-learning leveraging hand-crafted features (Aramaki et al., 2010; Chapman et al., 2019; Nikfarjam et al., 2015).

The 2018 n2c2 Adverse Drug Events and Medication Extraction in EHR data-set (Buchan et al.) and Medications and Adverse Drug Events from Electronic Health Records (MADE 1.0) (Jagannatha et al., 2019) are instances of ClinicalNLP shared-tasks focused on medication entity extrac-
Table 2: Dataset Characteristics.

| corpus | notes | Drug | Strength | Form | Frequency | Route | Dosage | Reason | ADE | Duration |
|--------|-------|------|----------|------|-----------|-------|--------|--------|-----|----------|
| training | 303 | 16225 | 6691 | 6651 | 6281 | 5476 | 4221 | 3855 | 959 | 592 |
| test     | 202 | 10575 | 4359 | 4230 | 4012 | 3513 | 2681 | 2545 | 625 | 378 |

3 Methods

3.1 Data and Pre-Processing

We use the 2018 n2c2 Adverse Drug Events and Medication Extraction (Track 2) data-set for our experiments. The data-set has a total of 505 clinical notes with nine medication-entities, as shown in Table 2. We convert these files into CoNLL 2000 BIO (Begin, Inside, Outside) format after pre-processing: split sentences into words, normalize numeric values, treat a subset of punctuation characters as word-boundary markers.

Relationship association for medication entities is complementary to our work and can be implemented either jointly or in a pipeline. Such a joint architecture utilizes the signals from the relations task to filter out unwanted medication entities. Wei et al. (2020) adopted such a joint-approach with a three-classifier ensemble achieving 52.95% in ADE extraction. Chen et al. (2020) also used a joint-architecture supplemented by UMLS (Bodenreider, 2004) concept lookups and unique modeling of temporal entities.

Dai et al. (2020) cascaded classifiers sequentially to widen the contextual information available for ADE identification. This model also facilitates improved identification when spans overlap. They evaluated ten pre-trained embedding models: half of them were based on MIMIC-III while the rest were general-purpose. Kim and Meystre (2020) uniquely leveraged SEARN (Daumé et al., 2009), a search-based prediction algorithm for its preference of precision over recall.

Our work is most similar to Miller et al. (2019); they demonstrate that strong medication extraction models can be constructed with minimal engineering using contextual embeddings. The main differences from above mentioned studies are the evaluation of a broader array of contemporary biomedical embeddings, detailed study of fine-tuning strategies, and augmentation methods for ADE extraction.
provided as part of the data-set to appraise our models using the test-set. We report the ‘Relaxed F1’ score per prevailing practice.

4 Model Selection Procedures

In Transfer Learning, the linguistic-information encoded by contextual embedding acts as a primary input to the downstream task layer (BiLSTM). Fine-tuning is generally accepted to be beneficial. However, it requires familiarity with the scripts / associated frameworks specific to the embedding and data-set adaptation.

4.1 BERT Embeddings

BERT models have close to a dozen layers (heads). Understanding the linguistic information encoded by these layers and their relative contribution to downstream tasks is an active research area (Liu et al., 2019b; Kovaleva et al., 2019). Flair uses the last four layers of the BERT models to generate embeddings by default.

1. Choice of Layers (4L vs All): The default setting of the end four transformer layers leads to sub-optimal performance (under-fitting) on the training set (Table 3, Row 1). Rather than choosing specific layers, we tried using all layers. This option generates a vast number of features (11 x 768), for the downstream task (Bi-LSTM), and causes training to run out-of-memory.

2. Scalar Mix (SM): As an alternate, we adopted Scalar Mix (Peters et al., 2018), a pooling mechanism on the layer-generated representations. Scalar Mix results in a reasonable number of features (768) and performs optimally (Row 2).

3. Mean-Pooling of sub-tokens (MP): BERT models uniquely use word-piece tokenization for out-of-vocabulary (OOV) words. Embeddings can be generated using first sub-token, or first and last sub-tokens, or using an aggregate (mean-pooling) of all sub-tokens. The latter provides best performance (Row 3).

These settings deliver optimal performance for the BERT-models.

4.2 Impact of adding Glove

Akbik et al. (2018) show that paired use of classic word embeddings (such as Glove) and contextual embeddings enhance NER task performance. Table 4 shows the impact of adding Glove. For the CB model, the noticeable gains were Reason (+1.00 F1) and ADE (+9.00 F1). For the FP model, ADE reduction (-2.00 F1) was offset by gains in Reason (+1.00 F1), Duration (+0.50 F1), and Drug (+0.40 F1). The EP model did not show any meaningful difference. We used the paired method for the rest of our experiments.

4.3 Flair Embedding Fine-Tuning

Language-model fine-tuning aims to improve the performance of Flair-PubMed contextual embeddings on speciality corpora. We performed fine-tuning for 10 epochs using the 4391 clinical notes from the i2b2/n2c2 data-sets. While all entities exhibited gains, the prominent gainers are shown in Table 5. We used this fine-tuned model for the rest of our experiments.
### 5 Discussion

Tables 6 and 7 show the overall performance of the various models. The prefixes (BB, CB, EP, FP) shows the contextual embedding used; and the suffix (Pr, Re, F1) shows the Precision, Recall, F1 metrics. The two highest F1 score for each entity are indicated via subscripts. The three most challenging entities are underlined. Table 8 shows the proportion of overlap between two entities. We use TP / (TP+FN) where TP is the number of ‘Gold’ entities identified correctly and FN is the number of mispredictions (‘Pred’). Smaller values indicate higher overlap.

#### 5.1 Error Analysis

1. **Drug**: BERT-models out-perform in the recognition of entities that are predominantly part of the clinical lexicon (e.g., *Drug* and *Form*) with CB model out-performing in both. We think that clinical note pre-training contributes to this out-performance. BERT-based models seem to misclassify *Drug* entities when special characters are involved. Consider the three sentences: ‘CONTRAINICATIONS FOR IV CONTRAST’, ‘C-SPINE WITHOUT CONTRAST’, ‘C-SPINE W/O CONTRAST’. ‘CONTRAST’\(^1\) is a gold *Drug* annotation. FP/EP models identify ‘CONTRAST’ in all the three sentences. BERT-models get the first and second one correctly while ignoring the last. Approximately 17 out of 31 references to ‘CONTRAST’ in the test-set are without special characters and hence recognized correctly by all models. The remaining ones are abbreviations such as ‘W/O’, ‘WW/O’, or terms such as ‘NON-CONTRAST’. These are ignored by the BERT models.

\(^1\) ‘contrast dye’ is given to a patient to accentuate structures in the CT Scan (Cedars-Sinai)

### Table 6: BB and CB Models

| Entity   | BB-Pr | BB-Re | BB-F1 | CB-Pr | CB-Re | CB-F1 |
|----------|-------|-------|-------|-------|-------|-------|
| Drug     | 95.24 | 94.64 | 94.94 | 95.76 | 94.23 | 95.00 |
| Strength | 97.94 | 97.95 | 97.85 | 97.30 | 97.99 | 97.64 |
| Duration | 88.86 | 80.16 | 84.28 | 80.32 | 81.48 | 85.67 |
| Route    | 95.59 | 94.93 | 95.26 | 95.69 | 94.79 | 95.24 |
| Form     | 96.83 | 94.70 | 95.10 | 95.72 | 94.90 | 95.96 |
| ADE      | 64.55 | 39.04 | 48.65 | 57.89 | 61.04 | 40.63 |
| Dosage   | 93.05 | 93.92 | 93.48 | 93.19 | 93.47 | 93.33 |
| Reason   | 77.00 | 59.08 | 66.84 | 80.71 | 75.52 | 67.17 |
| Frequency| 96.34 | 97.06 | 96.95 | 97.52 | 96.96 | 97.54 |
| Overall  | 94.32 | 91.34 | **92.84** | 94.85 | 90.91 | **92.85** |

### Table 7: EP and FP Models

| Gold | Pred | BB | CB | EP | FP |
|------|------|----|----|----|----|
| ADE  | Reason | 81.8% | 79.2% | 86.08% | 83.82% |
| Reason | ADE   | 97.32% | 96.6% | 97.97% | 97.09% |
| A/R  | Drug   | 98.8% | 98.48% | 98.60% | 99.19% |
| Form | Route  | 98.49% | 98.51% | 98.41% | 98.37% |
| Route | Form   | 98.43% | 98.57% | 98.66% | 98.43% |
| Dosage | Strength | 99.01% | 98.30% | 98.61% | 98.64% |
| Dosage | Frequency | 99.21% | 98.84% | 99.87% | 99.36% |
| Duration | Frequency | 96.80% | 96.55% | 96.58% | 96.01% |

### Table 8: Confusion Matrix

2. **Duration**: Having the fewest entities (378), *Duration* gets mislabeled maximally with *Frequency* and to a lesser degree with *Dosage*. *Henry et al. (2020)*’s observation that colloquial language use is a leading contributor to the confusion also implies the underlying context-sensitivity. In ‘CLOBETASOL ... x up to 2 weeks per month’, ‘2 weeks per month’ gets incorrectly tagged as *Frequency*. In Section 5.3 we show that ensembling FP model with any one of the other models delivers best overall *Duration* performance.

3. **Form and Route**: Unusual *Routes* (‘take one tab under your tongue’) were naturally ignored by all models. Commonly, the method of drug administration is used to describe the drug form also. In ‘Heparin 5,000 unit/mL Solution Sig: One (1) Injection TID (3 times a day)’, ‘Injection’ refers to the former and hence a *Route* while in ‘EGD with epinephrine injection and BICAP cautery’, it refers to the drug *Form*. Likewise, ‘infusion’ generates disagreement. BERT-models generally do well.

4. **Dosages and Strengths**: Dosages were mislabeled most commonly for *Strengths* (‘iron 0.5 ml per day’) by all models followed by *Frequency*. In ‘levophed @ 12 mcg/min’, the FP model identifies ‘mcg/min’ as ‘Strength’ (correctly) while other models identify ‘mcg/min’ as ‘Frequency’.
5. Each model uniquely detects several entities not detected by other models (Table 9). Consider the two sentences that occur next to each other in a clinical note: ‘could affect your Coumadin??????/warfarin dosage.’ ‘Coumadin (Warfarin) and diet:’. The former contains ‘?’ and ‘/’ inter-mixed with the entities. All models detect the entities in the second sentence. However, for the first sentence, the FP model identifies a single Drug entity Coumadin??????/warfarin while the others ignore it altogether.

6. ADE and Reason: FP model out-performed in ADE recognition (F1=53.08%) followed by the EP model (F1=50.73%). Although the top three models (FP, EP, BB) differ only marginally in Precision (0.6%) they exhibit significant divergence in Recall (+5.76%). There are three significant factors:

Mislabeling between ADE and Reason: CB model generates the highest number of mislabels (low recall) while EP does the best as shown in Table 8.

Mislabeling of ADE/Reason with Drug: In ‘Heme/onc was consulted regarding hemolysis and anticoagulation. ... Given her multiple indications for anticoagulation, decision was made to begin coumadin ...’, the first reference to ‘anticoagulation’ is a Drug gold annotation (‘blood thinners’) while the latter is a Reason (‘medical indication’). This example demonstrates the need for good contextual disambiguation. BB/FP models identify correctly. The EP model, ignores the former, and incorrectly identifies the latter as Drug. The CB model fails to identify both entities.

Incomplete word context: Often a Drug entity is needed to successfully infer the presence of an ADE or a Reason entity. However, it may occur in a subsequent sentence creating a challenge for the model. To verify this hypothesis, we evaluated model behavior by combining a sentence with one or more of its subsequent sentences. This is discussed in the next section.

5.2 Prediction-time Sentence Augmentation

We evaluated model behavior by combining a sentence with one or more of its subsequent sentences. For example, the ‘Look-ahead-1 strategy’, pairs a sentence with the one immediately following it. We progressively increased the pairing length up to a paragraph. Table 10 shows the ADE performance resulting from this augmentation strategy. Table 11 lists several examples (Drug entities are marked bold when they occur in the subsequent sentence).

1. Reason: ‘Hypothyroid’ is detected by augmentation due to the co-occurrence of ‘Syn-
Table 12: Ensembles

| Ensemble | ADE F1 | Reason F1 | Overall F1 |
|----------|--------|-----------|------------|
| FP+BB    | 55.21  | 69.28     | 92.80      |
| FP+CB    | 54.73  | 69.37     | 92.86      |
| FP+EP    | 55.77  | 69.60     | 92.82      |

Table 13: FP+EP Ensemble

| Entity    | Precision | Recall | F1  | F1 Δ |
|-----------|-----------|--------|-----|------|
| Drug      | 93.18     | 95.88  | 94.51| -0.24|
| Strength  | 97.56     | 98.30  | 97.93| -0.14|
| Duration  | 86.54     | 86.77  | 86.66| +1.01|
| Route     | 95.12     | 95.36  | 95.24| -0.14|
| Form      | 96.50     | 94.98  | 95.73| +0.16|
| ADE       | 58.90     | 52.96  | 55.77| +2.69|
| Dosage    | 92.26     | 94.67  | 93.45| +0.23|
| Reason    | 74.25     | 65.50  | 69.60| +1.03|
| Frequency | 96.27     | 97.86  | 97.06| -0.04|
| Overall   | 92.74     | 92.89  | 92.82| -0.10|

6 Limitations and Future Work

There are a few limitations in this study that we plan to address in future works:

1. We did not fine-tune BERT and ELMo-based embedding models. Doing so may alter the performance profile of these models. Hence, an apples-to-apples comparison between the models is not recommended.

2. Adoption of better tokenization methods (e.g., clinical text processing tools), and handling special-cases (such as abbreviations) may further enhance model robustness.

3. We also did not do an exhaustive survey of the available embeddings. There may be other more effective embeddings.

7 Conclusion

In this study, we presented strong performing transfer learning models for the extraction of medication entities using several biomedical contextual embeddings. Our experiments shed light on the strengths of the various embeddings: Flair-PubMed embedding out-performs in ADE extraction. BioBERT and ClinicalBERT embeddings outperform in recognition of Drug and Form medication entities. ELMo-PubMed embedding delivers competitive performance in all medication entities. We showed that sentence-augmentation and ensembling are viable strategies to enhance ADE performance. Our approach is free of hand-generated features and built using off-the-shelf neural models, default hyper-parameters, and training procedures. These factors decrease the development effort. A detailed analysis of embedding-specific factors contributing to mis-classification and inclusion of fine-tuning procedures are part of our ongoing work.
8 Acknowledgements

We thank the anonymous reviewers for their valuable suggestions and feedback. This work was supported by the biomedical AI groups of Amrita Technologies, Amritapuri, India and Amrita Institute of Medical Sciences, Kochi, India.

9 Availability of Data and Materials

1. The 2018 n2c2 ADE and Medication Extraction (Track 2) data-set is protected by Data Usage Agreement. It can be obtained from Harvard DBMI Portal.

2. The code and setup instructions used for the experiments in this paper is available from Git.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649.

Emily Alsentzer, John Murphy, William Boag, Weihung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Masuichi, Kayo Waki, and Kazuhiko Ohe. 2010. Extraction of adverse drug effects from clinical records. MedInfo, 160:739–743.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270.

Kevin Buchan, Kahyun Lee, Susanne Churchill, and Isaac Kohane. n2c2 2018—track 2: Adverse drug events and medication extraction in ehrs.

CDC. 2017. Adverse drug events in adults. https://www.cdc.gov/medicationsafety/adult_adversedrugevents.html, Last reviewed on 2017-10-17.

Cedars-Sinai. Ct scan of the abdomen. https://www.cedars-sinai.edu/Patients/Programs-and-Services/Imaging-Center/For-Patients/Exams-by-Procedure/CT-Scans/CT-Scan-of-the-Abdomen.aspx, Last Reviewed on.

Raghavendra Chalapathy, Capital Markets CRC, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional lstm-crf for clinical concept extraction. ClinicalNLP 2016, page 7.

Alec B Chapman, Kelly S Peterson, Patrick R Alba, Scott L DuVall, and Olga V Patterson. 2019. Detecting adverse drug events with rapidly trained classification models. Drug safety, 42(1):147–156.

Victor Chedid, Priya Vijayvargiya, and Michael Camilleri. 2018. Invited editorial: Advantages and limitations of faers in assessing adverse event reporting for eluxadoline. Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association, 16(3):336.

Jonathan H Chen, Mary K Goldstein, Steven M Asch, and Russ B Altman. 2016. Dynamically evolving clinical practices and implications for predicting medical decisions. In Biocomputing 2016: Proceedings of the Pacific Symposium, pages 195–206. World Scientific.

Long Chen, Yu Gu, Xin Ji, Zhiyong Sun, Haodan Li, Yuan Gao, and Yang Huang. 2020. Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning. Journal of the American Medical Informatics Association, 27(1):56–64.

Hong-Jie Dai, Chu-Hsien Su, and Chi-Shin Wu. 2020. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. Journal of the American Medical Informatics Association, 27(1):47–55.

Bharath Dandala, Venkata Joopudi, Ching-Huei Tsou, Jennifer J Liang, and Parthasarathy Suryanarayanan. 2020. Extraction of information related to drug safety surveillance from electronic health record notes: Joint modeling of entities and relations using knowledge-aware neural attentive models. JMIR medical informatics, 8(7):e18417.

Hal Daumé, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. Machine learning, 75(3):297–325.

Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How user intelligence is improving pubmed. Nature biotechnology, 36(10):937–945.
Feifan Liu, Abhyuday Jagannatha, and Hong Yu. 2020. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Keith B Hoffman, Andrea R Demakas, Mo Dimbil, Nicholas P Tatonetti, and Colin B Erdman. 2014. Stimulated reporting: the impact of us food and drug administration-issued alerts on the adverse event reporting system (faers). *Drug safety*, 37(11):971–980.

Ehtesham Iqbal, Robbie Mallah, Daniel Rhodes, Honghan Wu, Alvin Romero, Nynn Chang, Olubanke Dzahini, Chandra Pandey, Matthew Broadbent, Robert Stewart, et al. 2017. Adept, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. *PloS one*, 12(11):e0187121.

Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimiciii, a freely accessible critical care database. *Scientific data*, 3:160035.

Meizhi Ju, Nhung TH Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. An ensemble of neural models for nested adverse drug events and medication extraction with subwords. *Journal of the American Medical Informatics Association*, 27(1):22–30.

Youngjun Kim and Stéphane M Meystre. 2020. Ensemble method–based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association*, 27(1):31–38.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Feifan Liu, Abhyuday Jagannatha, and Hong Yu. 2019a. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019b. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL-HLT*, pages 1073–1094.

Timothy Miller, Alon Geva, and Dmitriy Digach. 2019. Extracting adverse drug event information with minimal engineering. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 22–27.

Thomas J Moore and Curt D Furberg. 2015. Electronic health data for postmarket surveillance: a vision not realized. *Drug safety*, 38(7):601–610.

Azadeh Nikfarjam, Abeed Sarker, Karen O’connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Yuqi Si, Jinqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Janet Sultana, Paola Cutroneo, and Gianluca Trifirò. 2013. Clinical and economic burden of adverse drug reactions. *Journal of pharmacology & pharmacotherapeutics*, 4(Suppl1):S73.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott I. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Qiang Wei, Zongcheng Ji, Zhiheng Li, Jingcheng Du, Jingqi Wang, Jun Xu, Yang Xiang, Firaq Tiryaki, Stephen Wu, Yaoyun Zhang, et al. 2020. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 27(1):13–21.

Hua Xu, Shane P Stenner, Son Doan, Kevin B Johnson, Lemuel R Waitman, and Joshua C Denny. 2010. Medex: a medication information extraction system...
for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

A Appendices

A.1 List of Hyper Parameters

1. LSTM: Single-Layer, Bi-Directional, 256 hidden states.

2. Locked dropout: 0.5.

3. Word dropout: 0.05.

4. SGD optimizer with initial learning rate: 0.1, annealing rate of 0.5, and patience of 3.

5. Batch Size: 16. For BERT experiments, we used a batch size of 8 to avoid GPU out-of-memory issues.

6. We train with both training and development data-set (train_with_dev=True).

7. All experiments were conducted on Google Colab GPU + High-RAM configuration.