HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

J Biomed Inform. 2020 October ; 110: 103552. doi:10.1016/j.jbi.2020.103552.

Adverse Drug Event Detection using Reason Assignments in FDA Drug Labels

Corey Sutphin[1], **Kahyun Lee**[2], **Antonio Jimeno Yepes, PhD**[3], **Özlem Uzuner, PhD**[2], **Bridget T. McInnes, PhD**[1]

[1]Virginia Commonwealth University, Richmond, VA, USA

[2]George Mason University, Fairfax, VA, USA

[3]IBM Research Australia, Melbourne, Australia

Abstract

Adverse drug events (ADEs) are unintended incidents that involve the taking of a medication. ADEs pose significant health and financial problems worldwide. Information about ADEs can inform health care and improve patient safety. However, much of this information is buried in narrative texts and needs to be extracted with Natural Language Processing techniques, in order to be useful to computerized methods. In this paper, we present three methods consisting of a Conditional Random Field (CRF), a bi-directional Long Short Term Memory Unit with a CRF (bi-LSTM+CRF), and several ensembles of the two for extracting ADEs and their reason from FDA Drug Labels. We map extracted ADEs to the Medical Dictionary for Regulatory Activities (MedDRA) terminology for normalization. We show that each of the CRF and bi-LSTM+CRF perform well on our task, but their combination is even stronger, achieving 0.93 F_1 in identification and 0.54 F_1 in normalization.

1 Introduction

Adverse drug events (ADEs) are undesirable incidents that often lead to hospitalization, and account for an estimated 12% of all emergency room visits[1]. The number of serious or life-threatening ADEs is increasing[2]. ADEs pose significant health and financial problems worldwide[3]. Advance knowledge of potential ADEs could help health care providers avoid these events but most of the information related to these events is documented in narrative texts that remain inaccessible to computerized methods. In addition, the FDA provides reporting systems for identification of unlabeled adverse events for drugs after they are released into the market, e.g. the FDA Adverse Event Reporting System (FAERS). Providing information about the labeled ADEs for each drug can help identify unknown adverse events. Identifying and normalizing these ADEs to MedDRA (Medical Dictionary for REgulatory Activities), the terminology used for FAERS, could speed up the identification of new ADEs.

Natural Language Processing (NLP) methods for named entity recognition can identify ADEs and put them into a structured format for access by computerized systems, enabling their incorporation into, for example, clinical decision support systems, and helping improve patient safety and quality of care[1]. Traditionally, Conditional Random Fields (CRFs) have

been shown to perform well for the task of entity recognition[4]. CRFs are probabilistic graphical machine learning algorithms. They are sequence learners which take previous annotations into consideration when determining the label of the current term. This property makes them well suited for entity recognition tasks.

Recently, deep learning has also been successfully applied to similar tasks. The deep learning approach studied in this paper, a bidirectional network with Long short-term memory units (LSTMs) and a neural CRF (bi-LSTM+CRF), has also been shown to perform well on entity recognition[5]. LSTMs take as their input not just the current input example, but also what they have previously seen in the past. Hence, they have two sources of input: their current state and their past states. This allows them to connect previous observations, such as words in a sentence, and learn dependencies of these words over arbitrarily long distances. In a bi-directional LSTMs, data is processed in both directions with two separate hidden layers, which are then fed forward into the same output layer. This allows the system to exploit context in both directions. As a final layer, this method contains a neural CRF to capture dependencies within the tagging sequence[5].

In this work, we adapt both CRFs and bi-LSTM+CRFs to the task of identifying ADEs and their reason from FDA drug labels. As an additional step in ADE extraction, we normalize extracted terms to MedDRA terminology using dictionaries and deep learning. We analyze the strengths of each of CRFs and bi-LSTM+CRFs at ADE extraction, and we propose an ensemble learner that can take advantage of strengths of both systems. We show that each of CRFs and bi-LSTM+CRF perform very well on ADE extraction. However, their combination shows the complementary nature of the two systems. These systems are accompanied by a dictionary-based normalizer that utilizes MetaMap to ground the ADE mentions to concepts in the Medical Dictionary for REgulatory Activities (MedDRA).

2 Background

2.1 Resources

Medical Dictionary for REgulatory Activities (MedDRA).—The Medical Dictionary for REgulatory Activities (MedDRA) is a terminology used to classify ADEs. It is used within a reporting analysis framework to quickly detect problems related to drug-based treatments. MedDRA terms are hierarchically organized. The System Organ Class (SOC) level includes the most general terms where the Low Level Terms (LLT) level includes more specific terminologies. Between SOC and LLT there are three intermediate levels: High Level Group Terms (HLGT), High Level Terms (HLT), and Preferred Terms (PT). In this work, we use MedDRA terms version 20.1.

Unified Medical Language System (UMLS).—The Unified Medical Language System (UMLS)[6] is a data warehouse containing three knowledge sources: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. The Metathesaurus, which contains approximately 2 million biomedical and clinical concepts identified using Concept Unique Identifiers (CUIs), is made of over 100 different terminologies that have been semi-automatically integrated into a single source. MedDRA is one such source in the UMLS. The Semantic Network consists of a set of broad subject categories called semantic types in

which each concept in the Metathesaurus is assigned one or more semantic type. For example, the semantic type of the concept C0206250 [Autonomic nerve] is Body Part, Organ, or Organ Component. Currently, there exist 135 semantic types in the Semantic Network. In this work, we use the Metathesaurus from UMLS version 2018AA to map the ADEs to MedDRA PTs.

MetaMap.—MetaMap[7] is a freely available concept mapping system that maps terms in biomedical text to CUIs in the UMLS. MetaMap also provides the semantic type information for each of the mappings. In this work, we use both CUIs and semantic types as features in our CRF-based ADE extraction system.

2.2 Related Work

A large body of work in text mining and biomedical NLP has been dedicated to extracting ADEs from electronic health records[8,9], scientific publications[10], social media[11–13] and FDA drug labels[14]. Here, we focus on approaches that extract ADEs from FDA drug labels. These approaches primarily fall into two categories: dictionary-based, and supervised machine learning approaches.

Many of the dictionary based approaches use information from the Unified Medical Language System Metathesaurus (UMLS). For example, the Side Effect Resource (SIDER)[15] and SciMiner[16] systems detect mentions of ADEs FDA Drug Labels based on synonyms derived from the MedDRA concepts in the UMLS[6]. The Structured Product Label Information Coder and ExtractoR (SPLICER)[17] system uses a series of regular expressions derived from a dictionary of ADEs extracted from the UMLS. Ly et al.[14] evaluated three NLP systems (ETHER, I2E, MetaMap) for extracting ADEs from drug labels and linking the terms to MedDRA Preferred Terms (PTs).

The supervised machine learning approaches utilize either Conditional Random Fields (CRFs) or Neural Networks (Deep Learning) approaches. For example, Zhou, et al[18] used a Convolutional Neural Network combined with Long Short Term Memory units with a CRF to identify side effects of 16 anti-Multiple Myeloma (MM) drugs from drug labels. The AutoMCExtractor[19] system used a set of CRF classifiers, trained on token, linguistic, and semantic features identified by cTAKES with a dictionary-based post-processing corrected boundary-detection errors of the CRF step.

Our work conducts a direct evaluation of a CRF-based approach (CRF) which use lexical, syntactic and semantic features identified by MetaMap to represent possible mentions, and a deep learning-based approach (bi-LSTM+CRF) that uses word and character embeddings to represent possible mentions. To combine the strengths of both approaches, we also evaluate three ensemble methods showing the complementary nature of the two entity detection approaches.

3 Methods

In this work, we develop NLP methods for identifying mentions of ADEs from narratives of FDA drug labels and we map each mention to MedDRA terminology. We study three

supervised machine learning methods for ADE extraction from FDA drug labels: The first is a CRF-based approach which uses lexical, syntactic and semantic features to represent possible mentions. The second is a deep learning-based approach which uses both lexical and character embeddings to represent possible mentions. Given these two systems, we present a third one that is an ensemble of the first two and complement them with two strategies for mapping the extracted entities to MedDRA terms, i.e., term normalization. The first normalization approach is dictionary-based while the second utilizes deep learning with an auto encoder/decoder network to learn the MedDRA Preferred Terms (PTs) associated with a term.

3.1 FDA Drug Label Dataset

The data consists of 100 annotated FDA drug labels (https://sites.mitre.org/adeeval), which include 15,562 mentions of the office of surveillance and epidemiology *(OSE) labeled ADEs* under 4 Reasons; 7,715 mentions of *non-OSE ADEs* under 10 Reasons; and 3,281 mentions of *Not ADE Candidates* under 4 Reasons. Each *OSE ADE* and *Non-OSE ADEs* are mapped to MedDRA terms version 20.1. The breakdown of the entities and their reasons are shown in Table 1.

3.2 Evaluation Metrics

We evaluate our methods using 5-fold cross validation and report the precision, recall and F_1 score for each entity. Precision is the ratio between correctly predicted mentions over the total set of predicted mentions for a specific entity; recall is the ratio of predicted mentions over the actual number of mentions; and F_1 is the harmonic mean between precision and recall. We also report the micro and macro averages over the entity types. Macro-average computes the metric independently for each class and takes the average treating all classes equally. Micro-average aggregates the contributions of all classes to compute the average metric. For multi-class classification, micro-averaging is preferable to understand the behaviour of a system if there is a class imbalance between the entity types.

3.3 ADE Extraction

We tackle ADE extraction using three methods. A CRF-based system, a bi-LSTM+CRF, and an ensemble of the two. The methods are trained over the Reasons associated with the ADE with the OSE and non-OSE labels extrapolated from the Reason annotations.

Conditional Random Field (CRF)-based ADE Extraction—For our CRF-based method, we adapted our python based entity recognition framework, medaCy (https://github.com/NLPatVCU/medaCy). MedaCy decomposes entity recognition into three sequential components: pre-processing, feature representation, and training/prediction. Each of these components are tuned and optimized for the particular entities in need of extraction. We used the following components:

Preprocessing.: The first preprocessing step removes non-ASCII characters from the documents; and converts XML formatting to BRAT annotation[20] format. For entities with discontinuous spans, we merge them into a single annotation. For example, the phrase ‘Bacterial, fungal, viral or protozoal infections’ contains four entities: ‘bacterial infections,

fungal infections. viral infections and protozoal infections, three of which are discontinuous in their span, i.e., there are four words between ‘bacterial’ and ‘infection’. Custom tokenization rules are created to decompose the text into atomic pieces for classification. This predefined list of rules allows for the inclusion of abbreviations and multi-word phrases to be treated as a single token (for instance the characters N.Y. would be kept together).

Features Representation.: The following features are utilized using the word ‘Bacterial’ from the phrase above as an example:

- Morphological Features
  - – Shape – capitalisation, punctuation, digits (e.g Xxxxxxxxx)
  - – Prefix/Suffix – the first and last three characters of the word (e.g. bac for prefix; ial for suffix)
- Lexical Features
  - – Token (e.g. bacterial)
  - – Features of the surrounding tokens within a window size of 3 (e.g. Features for fungal, viral and or)
- Syntactic Features
  - – Part of speech (POS) of the token (e.g. adjective (JJ))
- Semantic Features
  - – UMLS CUIs identified by Metamap (e.g. C0521009)
  - – Semantic Types (e.g. Qualitative Concept (qlco))
- Domain Specific Features
  - – ADE Lexicon membership

Training.: For training/prediction we use a linear chain CRF implementation called CRFsuite[21]. CRFs are probabilistic graphical machine learning algorithms. They are sequence learners that take previous annotations into consideration when determining the label of the current token making them well suited for the entity recognition task. The classifier is trained over the Reason labels described in the Data Section, and the OSE ADE and non-OSE ADE labels extrapolated from the Reason predictions.

Deep Learning-based ADE Extraction—For our deep learning-based method we developed a bidirectional Long Short Term Memory units that utilizes a CRF (bi-LSTM +CRF) for label sequence optimization. Our system consists of four components: preprocessing; feature representation, prediction/training; and postprocessing.

Preprocessing.: We process each drug label using SpaCy for sentence boundary detection and word tokenization. As in the CRF-based method, for entities with discontinuous spans, we merge them into a single annotation. We create a single entity from these annotations for

input to the system with the goal of first identifying the presence of discontinuous spans and then generating the separate entity annotations from system output at post-processing time.

Feature Representation.: Our feature representation consists of two layers: a character-embedding layer, and a token-embedding layer. We apply bi-LSTM for both the character-embedding and the token-embedding layers. We use character embeddings, token embeddings, and contextualized embeddings as features. For token embeddings, we use embeddings that are pre-trained on Wikipedia, PMC, and PubMed by Biomedical Natural Language Processing Lab[22]. We also utilize a context function which is pre-trained on the 1B Word Benchmark using ELMo[23]. We process the training data through the context function to extract contextualized embeddings. We concatenate character, token and contextualized embeddings before inputting the resulting vector into the label-prediction layer.

Training/Prediction.: Our training/prediction consists of two layers: a label-prediction layer, and a label-sequence-optimization layer[24]. We tune hyperparameters using five-fold cross-validation on the training set. The resulting hyperparameters are: character-embedding dimension of 50; token-embedding dimension of 200; contextualized-embedding dimension of 1,024; label-prediction dimension of 150; dropout probability of 0.5. We iterate the training with maximum epochs of 100. When we do not get any better micro averaged F_1 score than the previous best result within the last 10 epochs, we stop iterating. The classifier is trained over the Reason labels described in the Data Section, and the OSE ADE and non-OSE ADE labels extrapolated from the Reason predictions.

Ensemble Learner—Ensemble learning[25] is a paradigm where multiple machine learning algorithms are trained to solve the same problem. Ensemble methods try to construct a set of hypotheses and combine them, in contrast to traditional approaches which attempt to learn one hypothesis from a training data set.

In this work, we explore three approaches: union, intersection, and meta-learner. The union and intersection are simply the union or intersection mentions annotated by the CRF and bi-LSTM+CRF. The meta-learner approach combines the results of the CRF and bi-LSTM +CRF using a Naïve Bayes meta-learner to learn the label probability outputs of each ensemble member for each entity. We used SciKit Learn's[26] Gaussian Naive Bayes implementation for this work.

3.4 Normalization

To map the extracted ADEs to MedDRA PTs, we processed the text through MetaMap[7], which assigns Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) to biomedical text. From the UMLS CUI, we identify the MedDRA PT associated with that CUI. If one exists, we assign the MedDRA PT to the mention.

4 Results and Discussion

We evaluate the CRF, bi-LSTM+CRF, and their ensemble systems on the extraction of ADEs on FDA drug labels. As discussed in the Data Section, the data set contains *OSE*

ADEs, *non-OSE ADEs* and *Not ADE Candidates* where each mention is labeled with the *Reason* for the label. We evaluate the systems on both the ADE labels and Reasons. We evaluate our methods using 5-fold cross validation over the training data and report the precision, recall and F_1 scores.

4.1 Reason Results and Analysis

Table 2 shows the precision, recall and F_1 scores of the CRF and bi-LSTM+CRF trained on the ADE reasons using 5-fold cross validation. The table also shows the number of mentions for each of the entities in the data set. The results show that no one method obtained a higher F_1 score across all of the entities types, and overall the CRF obtained a higher macro-average while the bi-LSTM+CRF obtained a higher micro-average. Table 3 shows the number of disagreements for both systems.

Figure 1 shows the F_1 scores of the approaches sorted based on the number of instances. The results indicate that when there are few training examples the CRF approach generalizes better over the training data than the bi-LSTM+CRF approach. Figure 2 shows the average number of mis-labels across the different reasons for the CRF and bi-LSTM+CRF for each label type. The results indicate that both systems had difficulty identifying *from drug use*.

Analysis of the Reasons showed a high lexical variability of the entities. On average, there exists 2.05 possible Reasons for each lexical representation of an entity in the dataset. For example, the lexical phrase *antibodies to infliximab* was seen as an *AE only as instruction* 9 times and *from drug use* 8 times; its reasons depends on the context in which the term was used.

In addition, there are a number of mentions that can be both a Reason and Not a mention. For example, *vomiting* was labeled as a mention of a Reason in 103 instances while the system found an additional 19 mentions that were not labeled as a Reason. This high lexical ambiguity increases the difficulty of identifying the ADE mentions within the text.

4.2 ADE Results and Analysis

Table 4 shows the Precision, Recall and F_1 scores for the CRF, bi-LSTM+CRF, and ensemble approaches over the FDA ADE Labels. The results show that the bi-LSTM+CRF overall obtained higher precision, recall and F_1 score except for the precision of the *Not ADE Candidate* label. Compared to individual methods, the union ensemble shows a large recall improvement while the intersection ensemble provides a large precision improvement. The metalearner results are significantly lower. We believe this is due to the fact that the number of training examples was significantly lower for the *Non OSE ADE* and *Not ADE Candidate* labels. The *OSE ADE* label contained 16,000 instances, while the *Non OSE ADE* label 8,000 instances and the *Not ADE Candidate* contained only 8,000.

Table 5 shows the confusion matrix of the ADE Labels for both the CRF and bi-LSTM +CRF approaches. For both approaches, the largest confusion was between *Non-OSE ADE* and *OSE Labled ADE*. This makes sense because we would not expect lexical variability between an ADE that has been defined by the OSE and one that has been defined by non-OSE.

Merging these two labels increases the performance of both the CRF and bi-LSTM+CRF systems as shown in Table 6 increasing the F_1 scores to 0.88 & 0.93 respectively. These results can be compared to Ly et al.[14] results, showing that the methods proposed in this work largely improve the performance of previous work on a similar task.

4.3 Normalization Results and Analysis

In this section, we discuss our results assigning entities to MedDRA ids. Table 7 shows the precision, recall, F_1, true positive, false positive and false negatives for identifying the MedDRA PT for the OSE ADE lables identified by our system. The results show that using MetaMap provides a precision of (0.704), and a recall of (0.502).

Analysis of the results showed three areas that the system was unable to obtain the correct MedDRA PT:

1. MetaMap did not map the instance the correct CUI. For example, *Drug specific antibody* was mapped to 'C0443640:Specific antibody' rather than C4524162 which maps to the MedDRA ID 10080179.
2. A term may map to more than one CUI. For example, 'Hypertension' maps to more than one CUI C1963138 and C0020538 in which only C0020538 maps to 10020772.
3. A CUI may map to more than one MedDRA PT. For example, 'C0151740:Intracranial hypertension' maps to either the MedDRA ID '10011570:CSF pressure increased' or '10022773:Intracranial pressure increased '.

5 Conclusions

In this paper, we evaluated the performance of a CRF and a bi-LSTM+CRF approach to the identification of ADEs and their Reason from FDA drug labels. The results show that both methods obtain a high identification performance, even though the bi-LSTM+CRF approach improves over the CRF method. In the task of Reason identification, CRF performs better on those entities that have a low number of instances, while the bi-LSTM+CRF improves on those entities with a higher number of instances. The ensemble methods show an improvement of precision or recall at the cost of a reduced F1 indicating the complementary nature of the two systems.

The reported results improve previously results reported by Ly et al.[14] although a direct comparison can not be established. ADE reason identification could be improved for some of the categories, which might require additional annotations to be used for training the proposed machine learning methods. The developed solutions obtain a more than respectable performance on the proposed tasks and could be considered for the automatic identification of novel adverse drug events reported in FAERS or in other reporting database.

6 Future Work

For ADE extraction, we showed the complementary aspect of the CRF and bi-LSTM+CRF systems exploring three ensemble methods. In the future, we plan to develop more complex algorithms that take into account the probability of the systems annotations. We plan to expand out our meta-learner ensemble approach to include additional metadata from the two systems. In addition, there were 982 mentions in the data set that contained discontinuous spans. For example, the phrase 'Bacterial and fungal infections' contains two entities: one discontinuous spanned entity ('bacterial infections), and one continuous spanned entity (fungal infections). Currently, we merge them into a single annotation (e.g. 'bacterial and fungal infections', and 'fungal infections'). In the future, we would like to explore incorporating parsing information to take these entity types into account.

References

[1]. Banerjee R, Choi Y, Piyush G, Naik A, Ramakrishnan I. Automated suggestion of tests for identifying likelihood of adverse drug events. In: 2014 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2014. p. 170–175.

[2]. Deftereos SN, Andronis C, Friedla EJ, Persidis A, Persidis A. Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. 2011;3(3):323–334. [PubMed: 21416632]

[3]. Hristovski D, Kastrin A, Dinevski D, Burgun A, Ziberna L, Rindflesch TC. Using literature-based discovery toˇ explain adverse drug effects. Journal of medical systems. 2016;40(8):185. [PubMed: 27318993]

[4]. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001;.

[5]. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv preprint arXiv:160301360. 2016;.

[6]. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl 1):D267–D270. [PubMed: 14681409]

[7]. Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. 2001;.

[8]. Friedman C. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. In: Conference on Artificial Intelligence in Medicine in Europe. Springer; 2009. p. 1–5.

[9]. Haerian K, Varn D, Vaidya S, Ena L, Chase H, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clinical Pharmacology & Therapeutics. 2012;92(2):228–234. [PubMed: 22713699]

[10]. 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 safety. 2014;37(10):777–790. [PubMed: 25151493]

[11]. MacKinlay A, Aamer H, Jimeno Yepes A. Detection of Adverse Drug Reactions using Medical Named Entities on Twitter. In: AMIA Annual Symposium Proceedings. vol. 2017. American Medical Informatics Association; 2017. p. 1215.

[12]. Sarker A, Ginn R, Nikfarjam A, OConnor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: a review. Journal of biomedical informatics. 2015;54:202–212. [PubMed: 25720841]

[13]. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: a scoping review. Journal of medical Internet research. 2015;17(7).

[14]. Ly T, Pamer C, Dang O, Brajovic S, Haider S, Botsis T, et al. Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology. Journal of biomedical informatics. 2018;83:73–86. [PubMed: 29860093]

[15]. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic acids research. 2015;44(D1):D1075–D1079.

[16]. Hur J, Schuyler AD, States DJ, Feldman EL. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics. 2009;25(6):838–840. [PubMed: 19188191]

[17]. Duke JD, Friedlin J. ADESSA: a real-time decision support service for delivery of semantically coded adverse drug event data. In: AMIA Annual symposium proceedings. vol. 2010. American Medical Informatics Association; 2010. p. 177.

[18]. Zhou K, Zhang S, Meng X, Luo Q, Wang Y, Ding K, et al. CRF-LSTM Text Mining Method Unveiling the Pharmacological Mechanism of Off-target Side Effect of Anti-Multiple Myeloma Drugs. In: Proceedings of the BioNLP 2018 workshop; 2018. p. 166–171.

[19]. Li Q, Deleger L, Lingren T, Zhai H, Kaiser M, Stoutenborough L, et al. Mining FDA drug labels for medical conditions. BMC medical informatics and decision making. 2013;13(1):53. [PubMed: 23617267]

[20]. Stenetorp P, Pyysalo S, Topic G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted´ text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; 2012. p. 102–107.

[21]. Okazaki N. CRFsuite: a fast implementation of Conditional Random Fields (CRFs); 2007. Available from: http://www.chokkan.org/software/crfsuite/.

[22]. Moen S, Ananiadou TSS. Distributional semantics resources for biomedical text processing. Proceedings of LBM. 2013;p. 39–44.

[23]. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv preprint arXiv:180205365. 2018;.

[24]. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017;24(3):596–606. [PubMed: 28040687]

[25]. Dietterich TG, et al. Ensemble learning. The handbook of brain theory and neural networks. 2002;2:110–125.

[26]. Bressert E. SciPy and NumPy: An Overview for Developers. 1st ed. 1005 Gravenstein Highway North Sebastopol, CA 95472: O'Reilly Media; 2013.

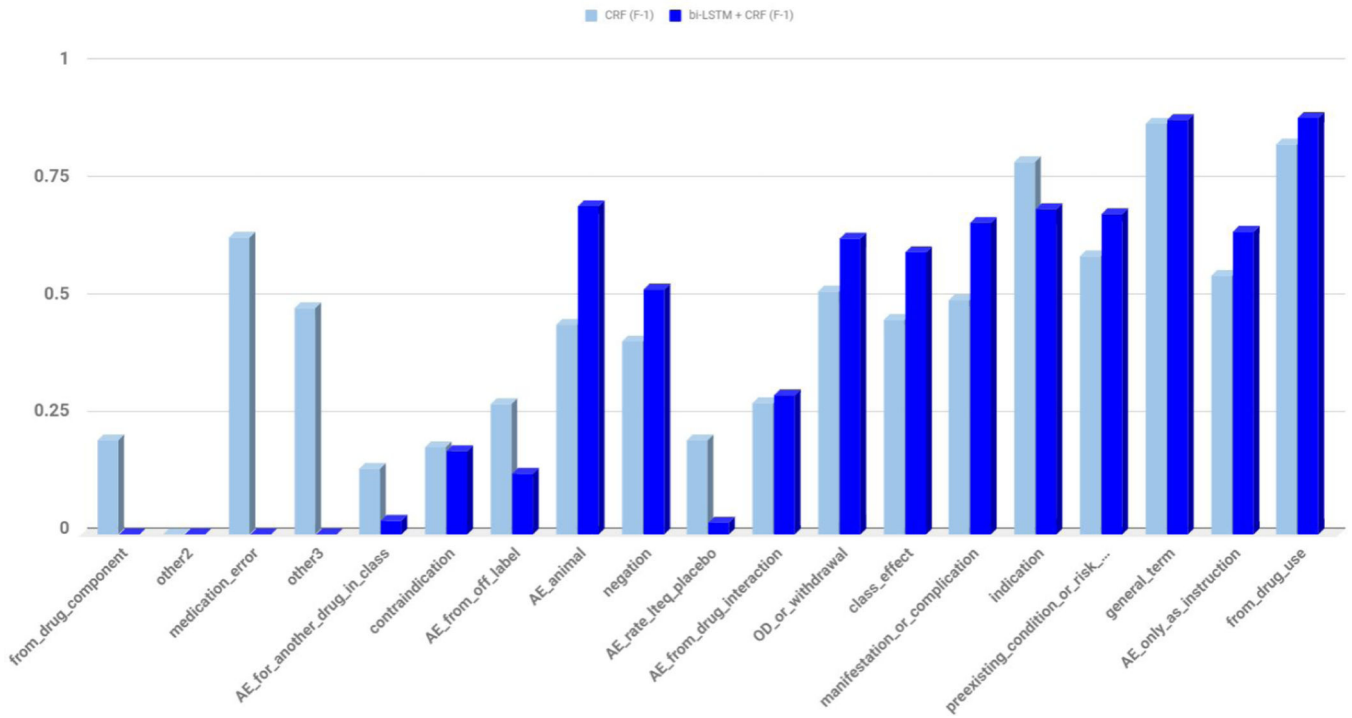

Figure 1:
F_1 scores sorted based on number of instances in the data set.

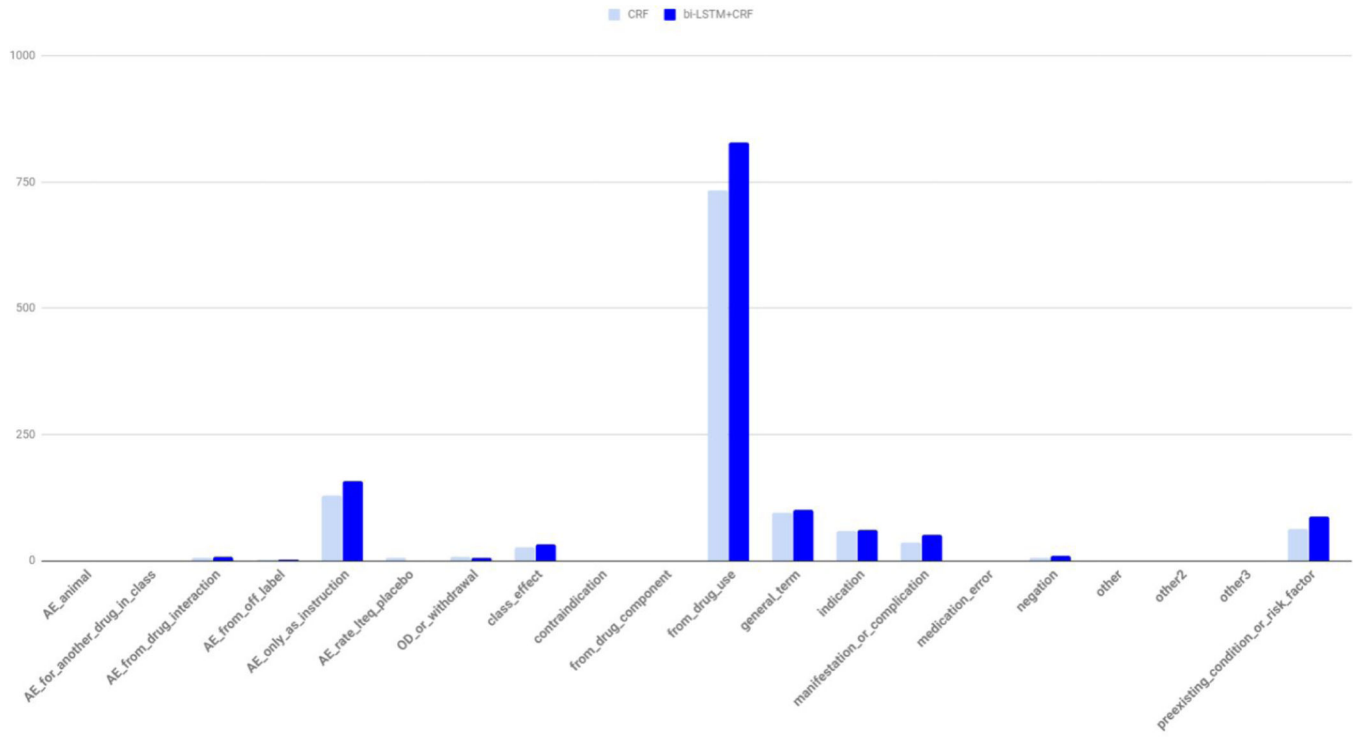

Figure 2:
The average number of mis-labels across the CRF and bi-LSTM+CRF for each entity type

Table 1:

FDA Drug Label Data Set

Label	Reason	# Instances	Description
OSE ADE	from drug use	14936	ADE associated with the use of the drug.
	from drug component	1	ADE associated with an inactive ingredient.
	class effect	587	Effect associated with the drug class.
	medication error	7	Preventable event that may lead to inappropriate drug use.
Non-OSE ADE	manifestation or complication	760	Signs, Symptoms or changes in lab results related to the ADE.
	ADE rate lteq placebo	119	ADE with incident rate equal or lower than placebo.
	ADE animal	76	ADE from animal data.
	ADE from drug interaction	140	ADE from drug-drug interaction.
	general term	1952	Non specific text use to introduce ADEs.
	ADE from off label	47	ADE associated with off-label use.
	ADE only as instruction	2824	ADE mentioned in instructions.
	ADE for another drug in class	17	ADE related ot another drug class.
	OD or withdrawal	157	ADE associated with discontinuing medication.
	negation	110	ADE whose presence is negated.
	other	2	Another reason for disinterest.
Not ADE Candidate	indication	1230	Clinical reason for taking the drug.
	contraindication	25	Clinical symptom for which the use of the drug would not be appropriate.
	preexisting condition or risk factor	1365	Clinical symptom of pre-existing condition.
	other	10	Another reasons the mention is not an ADE.

Table 2:

Overall CRF & bi-LSTM+CRF Precision, Recall and F_1 Results over the Reasons

Label	Reason	# instances	CRF			bi-LSTM+CRF		
			Precision	Recall	F_1	Precision	Recall	F_1
OSE ADE	from drug use	14936	0.8247	0.8357	0.8302	0.8561	0.9189	**0.8864**
	from drug component	1	1.000	0.1111	**0.200**	0.000	0.0000	0.0000
	class effect	587	0.5366	0.3982	0.4572	0.686	0.5336	**0.6003**
	medication error	7	0.8571	0.500	**0.6316**	0.000	0.0000	0.0000
Non-OSE ADE	manifestation or complication	760	0.5908	0.4326	0.4994	0.6830	0.6451	**0.6635**
	ADE rate lteq placebo	119	0.3445	0.1419	**0.2010**	0.1667	0.0138	0.0256
	ADE animal	76	0.8026	0.3096	0.4469	0.7487	0.6564	**0.6995**
	ADE from drug interaction	140	0.4500	0.2032	0.2800	0.4294	0.2258	**0.2960**
	general term	1952	0.9109	0.8411	0.8746	0.8896	0.8749	**0.8822**
	ADE from off label	47	0.3830	0.2169	**0.2769**	0.2000	0.0964	0.1301
	ADE only as instruction	2824	0.5857	0.5187	0.5501	0.6366	0.6538	**0.6451**
	ADE for another drug in class	17	0.2941	0.0909	**0.1389**	0.0625	0.0182	0.0282
	OD or withdrawal	157	0.5924	0.4604	0.5181	0.9238	0.4778	**0.6299**
	negation	110	0.6364	0.3030	0.4106	0.5459	0.4979	**0.5208**
	other2	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Not ADE Candidate	indication	1230	0.8472	0.7443	**0.7924**	0.7157	0.6699	0.6921
	contraindication	25	0.3200	0.1311	0.1860	0.3889	0.1148	**0.1772**
	preexisting condition or risk factor	1365	0.6821	0.5224	0.5917	0.6794	0.6876	**0.6835**
	other3	10	0.7000	0.3684	**0.4828**	0.0000	0.0000	0.0000
AVERAGE	**macro**		**0.5679**	**0.3565**	**0.4184**	0.4306	0.3542	0.3780
	micro		0.7741	0.7109	0.7411	**0.7946**	**0.7983**	**0.7965**

Table 3:

Disagreements between CRF & bi-LSTM+CRF

CRF	bi-LSTM+CRF	# mentions
manifestation or complication	from drug use	229
preexisting condition or risk factor	from drug use	147
AE only as instruction	from drug use	44
class effect	from drug use	102
preexisting condition or risk factor	None	425
from drug use	manifestation or complication	151
None	AE only as instruction	378
from drug use	class effect	154
general term	None	271
from drug use	AE only as instruction	415

Table 4:

Precision, Recall and F_1 Lenient Results over the Labels

Method	Label	Precision	Recall	F_1
CRF	OSE ADE	0.8376	0.8361	0.8368
	Non OSE ADE	0.7582	0.6102	0.6762
	Not ADE Candidate	**0.8224**	0.6627	0.7340
bi-LSTM+CRF	OSE ADE	**0.8738**	**0.9197**	**0.8962**
	Non OSE ADE	**0.7996**	**0.7308**	**0.7637**
	Not ADE Candidate	0.8001	**0.7648**	**0.7820**
UNION ENSEMBLE	OSE ADE	0.8392	**0.9386**	**0.8861**
	Non OSE ADE	0.7675	**0.7611**	**0.7643**
	Not ADE Candidate	0.7802	**0.8225**	**0.8008**
INTERSECT ENSEMBLE	OSE ADE	**0.8980**	0.8384	0.8672
	Non OSE ADE	**0.8447**	0.6170	0.7131
	Not ADE Candidate	**0.8526**	0.6357	0.7283
META-LEARNER ENSEMBLE	OSE ADE	0.7104	0.8426	0.7708
	Non OSE AE	0.4280	0.4677	0.4470
	Not ADE Candidate	0.6767	0.5294	0.5940

Table 5:

Confusion Matrix over the Labels

CRF Results	**OSE ADE**	**Non-OSE ADE**	**Not ADE Candidate**
OSE ADE	12992	1285	198
Non-OSE ADE	733	4693	179
Not ADE Candidate	74	111	2158
bi-LSTM+CRF Results	**OSE ADE**	**Non-OSE ADE**	**Not ADE Candidate**
OSE ADE	14962	1296	151
Non-OSE ADE	581	5799	230
Not ADE Candidate	97	180	2564

Table 6:

ADE Results with Merged OSE ADE and Non-OSE ADE

Method	Precision	Recall	F_1 Score
CRF	0.9078	0.8480	0.8768
bi-LSTM+CRF	**0.9287**	**0.9300**	**0.9293**
UNION ENSEMBLE	0.8935	**0.9560**	**0.9237**
INTERSECT ENSEMBLE	**0.9520**	0.8196	0.8808
META-LEARNER ENSEMBLE	0.8835	0.9210	0.9019

Table 7:

Normalization Results

	Label	Precision	Recall	F_1	TP	FP	FN
Normalization Results on Predicted Mentions	OSE ADE	0.704	0.502	0.586	8243	3462	6843
	Non OSE ADE	0.523	0.231	0.321	1881	1710	5592
	Combined	0.736	0.459	0.566	11267	4029	11050
Normalization Results on Predicted Mentions	OSE ADE	0.841	0.588	0.692	9653	1822	5100
	Non OSE ADE	0.747	0.406	0.526	3296	1115	3794
	Combined	0.815	0.528	0.641	12949	2937	8886