UTH_CCB System for Adverse Drug Reaction Extraction from Drug Labels at TAC-ADR 2017

Jun Xu*, **Hee-Jin Lee***, **Zongcheng Ji***, **Jingqi Wang, Qiang Wei,** and **Hua Xu**
The University of Texas Health Science Center at Houston
Houston, TX
{Jun.Xu,Hee.Jin.Lee,Zongcheng.Ji,Jingqi.Wang,Qang.Wei,Hua.Xu}@uth.tmc.edu

Abstract

This paper describes the end-to-end system developed by the University of Texas Health Science Center at Houston, Center for Computational Biomedicine (UTHeath-CCB) team for the 2017 TAC track on "Adverse Drug Reaction Extraction from Drug Labels". Our system primarily uses machine learning and deep learning based approaches and it achieved competitive results on all four tasks in the challenge. The highest scores of our system on the test set are: 82.48% (micro-F1), 49.00% (micro-F1), 82.19% (macro-F1) and 85.33% (macro-F1) for tasks 1, 2, 3, and 4 respectively.

1 Introduction

Knowledge bases containing drugs and their adverse reactions are important for clinical research and practice. However, much of the drug and its adverse drug reaction (ADR) information are only available in narrative formats, such as drug labels and biomedical literature. Manual curation of these textual resources is often costly and time-consuming, making it difficult to keep the information up-to-date. Many text-mining tools such as CD-REST[1], SPLICER[2] have been developed to automatically extract information about drugs, ADRs, and the relations among them from text. Further, other studies describing systems that extract drug and ADR information from social media[3] and electronic health records also exist.[4, 5]

In 2017, the Text Analysis Conference organized a shared task entitled "Adverse Drug Reaction Extraction from Drug Labels", which aimed to examine current text mining methods on extracting ADR information from drug labels and normalizing them to concepts in MedDRA. The shared task consisted of 4 sub-tasks: 1) Task 1 – Extract mentions of *AdverseReactions* and modifier concepts (i.e., *Severity*, *Factor*, *DrugClass*, *Negation*, and *Animal)*; 2) Task 2 – Identify the relations between *AdverseReactions* and their modifier concepts (i.e., *Negated*, *Hypothetical*, and *Effect*); 3) Task 3 – Identify positive *AdverseReaction* mentions in the labels; and 4) Task 4 – Map recognized positive AdverseReaction to *MedDRA PT*(s) and *LLT*(s). In this paper, we describe our approaches and results for all the four tasks.

2 Methods

2.1 Datasets and pre-processing

For this track, the organizers prepared three datasets: 1) a training set of 101 annotated drug labels; 2) a development set of 2,208 un-annotated drug labels; and 3) a test set of 99 annotated drug labels, which were selected from the un-annotated development data set. We developed our models and optimized their parameters using the training set.

We used the CLAMP (Clinical Language Annotation, Modeling, and Processing) Toolkit[6] for pre-processing of drug label documents, including sentence boundary detection, tokenization, and POS tagging. We also utilized CLAMP APIs for entity recognition and normalization.

* Contributed equally to this work.

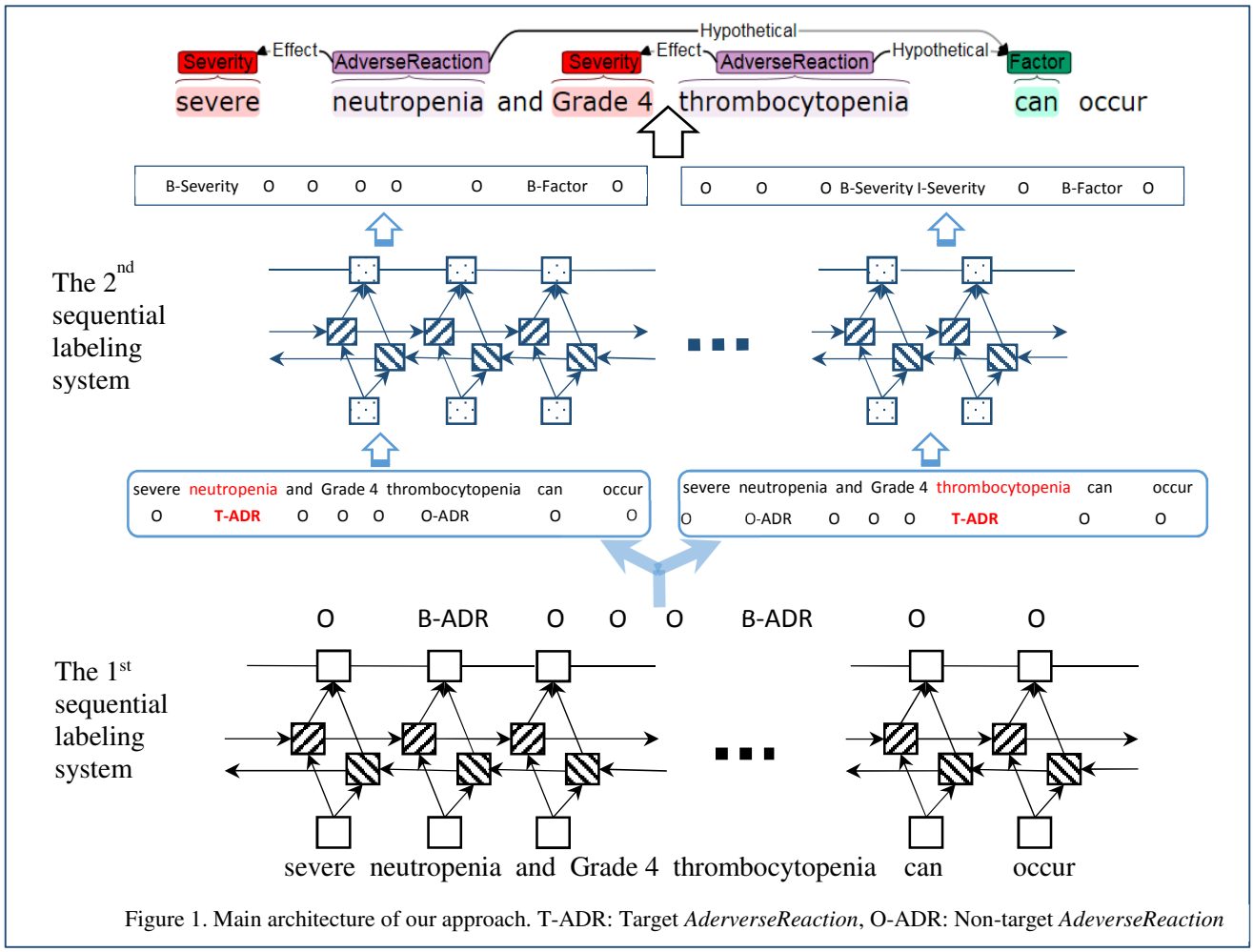

Figure 1. Main architecture of our approach. T-ADR: Target *AderverseReaction*, O-ADR: Non-target *AdeverseReaction*

2.2 Task 1&2 – Extract AdverseReactions, Related Concept Mentions, and their Relations

Task 1 is a typical named entity recognition (NER) task, and Task 2 is a typical relation extraction task. However, the one caveat is that, for Task 1, a mention of a modifier concept is to be extracted only if the modifier concept is associated with at least one *AdverseReaction*. When a modifier concept such as *Animal*, *DrugClass*, *Negation*, etc, is not associated with an *AdverseReaction*, it is not annotated in the gold standard. This meant that we could not directly train a machine learning-based NER model for detecting all modifiers, since all modifiers are not annotated. This also brings challenges to the conventional relation extraction approach that requires entity recognition first before generating candidate relation pairs for training. Since we did not have annotations for modifiers that are not associated with any of *AdverseReactions*, we would miss negative examples of relations that are required to train an effective relation classifier.

To address the above issues, we propose a cascaded sequence labeling approach that can address Task 1&2 at the same time. Figure 1 shows the architecture of our approach on Task 1&2. It cascades two sequential labeling systems. The first sequence labeling system recognizes all *AdverseReaction* mentions in one sentence, just like a typical NER system. The second sequence labeler extracts mentions of modifier concepts that are associated with an *AdverseReaction,* and classifies the type of the relation between the modifier concept and the *AdverseReaction.* Note that this labeler identifies both the modifier concepts and their relations to *AdverseReaction* in one-step.

This one-step approach is done by a new transformation schema. Given a sentence and *AdverseReactions* in the sentence that is identified by the first sequence labeler, we generate a sample of labeled sequence for each *AdverseReaction* mention in the sentence. That is, if more than one *AdverseReaction* mention is identified in a sentence, we generate multiple labeled sequence samples. For instance, in Figure 1, the sentence has two *AdverseReaction* mentions predicted by the first labeler, "neutropenia" and "thrombocytopenia". Thus, we generate two labeled sentence samples using each *AdverseReaction* mention as target (i.e., one with 'neutropenia' as target *AdverseReaction*, and another with 'thrombocytopenia' as target *AdverseReaction*). For each sample, we only label the mentions of modifier concepts that are associated with the target *AdverseReaction* using B or I together with the modifier type. For example, in the sample generated for "neutropenia", only "severe" and "can" are labeled as modifiers. Other modifiers like "Grade 4" which are not associated with "neutropenia", will not be annotated as a modifier. Then, we provide the position information of the target *AdverseReaction* mention to the second sequential labeler during the training and prediction.

Commonly, a named entity mention is a continuous string of text, such as "grade 3". In contrast, a discontinuous mention is represented in a discontinuous string of text. For instance, the phrase "grade 3 and 4" contains a continuous mention of type *Severity*, "grade 3", and a discontinuous mention of type *Severity*, "grade … 4". Since about 7% of the mentions in the gold standard training set are discontinuous, we propose a method that uses "fabricated" continuous mentions and rule-based splitting, to handle such discontinuous mentions.

Figure 2 shows the method. The method is composed of three steps. First, before training a seqeunce labeler, discontinuous mentions are merged with overlapped mentions to generate "fabricated" continuou mentions. For example, for the phrase "grade 3 and 4", the continuous mention "grade 3" and the discontinuous mention "grade ... 4" are merged into a single fabricated continuous mention "grade 3 and 4". Second, a sequence labeler is trained on a training set that contains both fabricated continuous mentions and original continuous mentions. In the prediction time, the sequence labeler will predict only continuous mentions for both original and fabricated mentions. Lastly, regular expression rules and dictionary-based rules are employed to detect fabricated continuous mentions and split them into continuous and/or discontinuous mentions. The rules are developed based on the observation on the training dataset. Given mentions predicted by the sequence labeler, any mention that has more than 4 tokens and contain any of 'and', 'or', '/', ',', or '(', is regarded as a fabricated continuous mention, and thus processed by the regular expresson rules or the dictionary-based rules to be split into continuous and discontinuous mentions.

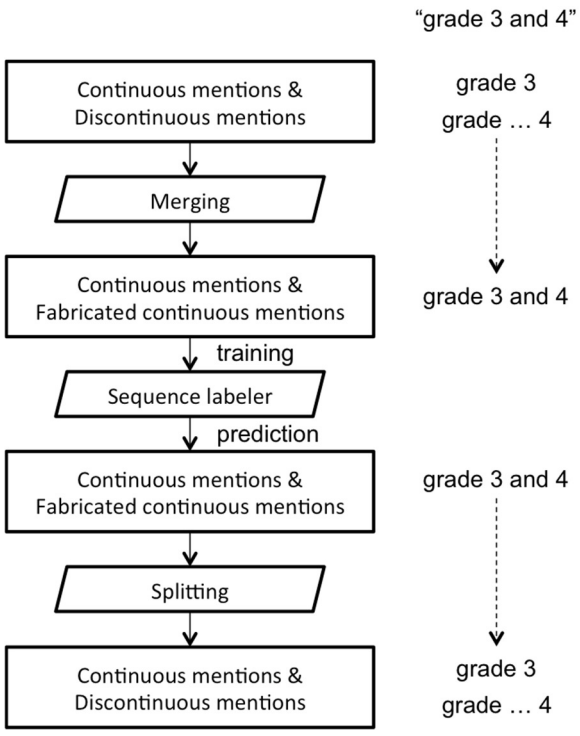

Figure 2. Method for discontinuous mention handling.

The regular expression rules use regex groups. For instance, to split "grade 3 and 4", we use the following regular expression rule:

((grade|stage)\s+\d)\s(?:and|or|\-|\/)\s*(\d)* →*1|2+3*

Using the above rule, the text string matched to the first regex group (i.e. "grade 3") will be identified as a continuous mention, and the text strings matched to the second and the third regex

groups (i.e., "grade" and "4", respectively) will be combined into a discontinuous mention (i.e., "grade … 4").

The dictionary-based rules use a dictionary of common phrase pairs, such as '<infections, viral>', '<infections, protozoal>', and '<increase in, AST>'. When there are two non-overlapping phrases in a fabricated continuous mention that matches a pair in the dictionary, the phrases are identified as a discontinuous mention. For example, from a fabricated continuous mention "viral, or protozoal infections", "viral … infections" is identified as a discontinuous mention along with a continuous mention "protozoal infections".

For sequence labeling, we use the LSTM-CRF[7] recurrent neural network as the method. The default LSTM-CRF is augmented with an additional embedding layer, to incorporate dictionary matching results produced from dictionaries of MedDRA terms and common modifiers. For the initial values of word embeddings, we employ Word2vec[8] and trained word embedings on the training and the development set.

2.3 Task 3 – Identify Positive AdverseReactions

Task 3 is to identify all the positive *AdverseReaction* mentions in the drug labels. To perform this task, we filter out negative *AdverseReactions* based on heuristic rules. An AdverseReaction mention is negative if one of the following two conditions hold: 1) the *AdverseReaction* is negated; 2) the *AdverseReaction* is related by a *Hypothetical* relation to a *DrugClass* or *Animal*. The remaining *AdverseReactions* are selected as positive *AdverseReactions*.

2.4 Task 4 – Normalize AdverseReactions to MedDRA PT(s) and LLT(s)

Task 4 is to normalize each positive *AdverseReaction* mention to an entry in MedDRA v18.1. We use a learning to rank technique to perform the normalization task. Formally, for a given mention m, we select the best MedDRA term d^* with the highest-ranking score from the repository.

$$d^* = \arg\max_d \mathbf{w}^{\mathrm{T}} \cdot \Phi(m, d)$$

where $\Phi(m, d)$ is the matching features between the mention m and a candidate MedDRA term d, and $\mathbf{w}$ is the corresponding feature weights.

More specifically, we first employ the BM25 model provided by Lucene to retrieve the top 10 candidate MedDRA terms for an *AdverseReaction* mention. Then, for each pair of an *AdverseReaction* mention and a candidate MedDRA term, we calculate three scores as matching features: BM25 ranking score, Jaccard similarity score and translation-based ranking score[9]. Finally, we employ the linear RankSVM[10], one of the widely-used methods for learning to rank, to assign a final ranking score to each candidate MedDRA term. The top ranked MedDRA term for each *AdverseReaction* mention is then chosen as the MedDRA normalization for the mention.

2.5 Submissions and Evaluation

We submitted three different runs:

- **Run 1**: We discarded all discontinuous *AdverseReaction* mentions in both training and prediction stages, to get a higher precision on task 1.
- **Run 2**: We merged all discontinuous mentions into continuous ones. We first identified all the continuous mentions and then split them back into discontinuous ones with rules as described in Section 2.2.
- **Run 3**: We combined the outputs of Run 1 and Run 2. All continuous *AdverseReactions* mentions and their modifier concepts from Run 1, and all discontinuous *AdverseReactions* mentions and their modifier concepts from Run 2 were merged into Run 3.

The evaluation metrics include Precision (P), Recall (R) and F1-measure (F1). Micro-averaged F1 is used as the primary metric for Task 1 and 2 and macro-averaged F1 is used as the primary metric for Task 3 and 4. For more details, please refer to the task description paper or the task website[†].

[†] https://bionlp.nlm.nih.gov/tac2017adversereactions/

3 Results and Discussion

Table 1. The performances of the three runs of our system on Task 1(*: Primary metric).

#Run		P	R	F1
1	+type	83.78	79.74	81.71*
	-type	83.83	79.79	81.76
2	+type	80.22	84.40	82.26*
	-type	80.25	84.42	82.28
3	+type	82.54	82.42	**82.48***
	-type	82.59	82.48	82.54

Table 2. The performances of the three runs of our system on Task 2(*: Primary metric).

#Run		P	R	F1
1	Full(+type)	51.67	44.45	47.79*
	Full(-type)	52.20	44.91	48.28
	Binary(+type)	55.51	50.86	53.09
	Binary(-type)	55.99	51.30	53.55
2	Full(+type)	46.24	48.32	47.26*
	Full(-type)	46.57	48.66	47.59
	Binary(+type)	50.19	56.73	53.26
	Binary(-type)	50.52	57.10	53.61
3	Full(+type)	50.24	47.82	**49.00***
	Full(-type)	50.72	48.28	49.47
	Binary(+type)	53.92	54.49	54.21
	Binary(-type)	54.36	54.93	54.64

Table 1 and 2 show the overall performance of the three runs of our system on Task 1 and 2, respectively. As we expected, Run1 achieved the highest precision on both tasks. Run 2 achieved the highest recall since the system recognize discontinuous *AdverseReactions*. Run 3 achieved the highest F1.

Table 3. The performances of the three runs of our system on Task 3(*: Primary metric).

#Run		P	R	F1
1	Micro-	82.83	81.76	82.29
	Macro-	82.61	81.88	81.65*
2	Micro-	79.68	85.57	82.52
	Macro-	78.77	85.62	81.39*
3	Micro-	80.97	84.87	82.87
	Macro-	80.69	85.05	**82.19***

Table 4. The performances of the three runs of our system on Task 4(*: Primary metric).

#Run		P	R	F1
1	Micro-	85.00	87.75	86.35
	Macro-	84.04	86.67	84.79*
2	Micro-	82.42	90.78	86.40
	Macro-	80.83	89.90	84.53*
3	Micro-	84.17	89.84	86.91
	Macro-	83.02	89.06	**85.33***

Table 3 and 4 show the overall performance of three runs of our system on task 3 and 4, respectively. Run 3 achieved the highest F1. Although the performance on relation classification (Task 2) was not high, we still got high performances on Task 3 and Task 4. These results also demonstrate the effectiveness of the end-to-end system we developed.

4 Conclusion and Future Work

In this paper, we describe our participation in the TAC 2017 ADR challenge – "Adverse Drug Reaction Extraction from Drug Labels". Our system took part in all the four sub-tasks. Our results show that it is feasible to extract adverse drug reactions from drug labels using machine-learning methods with high performance.

Acknowledgments

This study was supported by grants from the NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282.

References

[1]. Xu J, Wu Y, Zhang Y, Wang J, Lee H-J, Xu H. CD-REST: a system for extracting chemical-induced disease relation in literature. Database. 2016;**2016**:baw036-baw.

[2]. Duke JD, Friedlin J. ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. AMIA Annual Symposium Proceedings. 2010 11/13;**2010**:177-81.

[3]. Freifeld CC, Brownstein JS, Menone CM, et al. Digital Drug Safety Surveillance: Monitoring Pharmaceutical Products in Twitter. Drug Safety. 2014 04/29;**37**(5):343-50.

[4]. Haerian K, Varn D, Vaidya S, Ena L, Chase HS, Friedman C. Detection of Pharmacovigilance-Related adverse Events Using Electronic Health Records and automated Methods. Clinical pharmacology and therapeutics. 2012 06/20;**92**(2):228-34.

[5]. Wang X, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. Journal of the American Medical Informatics Association : JAMIA. 2009;**16**(3):328-37.

[6]. Soysal E, Wang J, Jiang M, et al. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. JAMIA. 2017.

[7]. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. NAACL-HLT; 2016; San Diego, US: Association for Computational Linguistics; 2016. p. 260-70.

[8]. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Lake Tahoe, Nevada: Curran Associates Inc.; 2013. p. 3111-9.

[9]. Ji Z, Lu Z, Li H. An Information Retrieval Approach to Short Text Conversation. eprint arXiv:1408.6988: ARXIV; 2014.

[10]. Lee C-P, Lin C-J. Large-Scale Linear RankSVM. Neural Computation. 2014;**26**(4):781-817.