A Text Analysis Conference (TAC) 2017 Research Paper

Extracting and Normalizing Adverse Drug Reactions from Drug Labels

Carson Tao[1], Kahyun Lee[1], Michele Filannino[1,2], Kevin Buchan[1], Kathy Lee[3], Tilak Arora[3], Joey Liu[3], Oladimeji Farri[3], Özlem Uzuner[4]

[1] State University of New York at Albany, Albany, NY, USA
[2] Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Philips Research North America, Cambridge, MA, USA
[4] George Mason University, Fairfax, VA, USA

Abstract

Drug labels provide descriptive information on the pharmaceutical properties of medications. These properties cover different types of prescription information (e.g., medication name, frequency, duration, and dosage) and drug-related safety concerns (e.g., potential adverse drug reactions). An adverse drug reaction (ADR) describes an undesirable effect associated with the administration of a particular medication. We propose a system that automatically extracts ADRs with their related mentions (e.g., drug classes, ADR severities), separates negated or hypothetical ADRs from others, and normalizes non-negated ADRs into MedDRA Preferred Terms (PTs) and Lowest Level Terms (LLTs). When evaluated on the official test set from the 2017 TAC ADR shared tasks, our best system achieved a micro-average F1 score of 0.702 on the extraction of ADRs and related mentions (exact match). The relation extraction module achieved an F1 score of 0.310. Finally, the identification of positive ADRs (Task 3) and their normalization (Task 4) achieved F1 scores of 0.703 and 0.780 respectively.

1. Introduction

Drug labels provide descriptive information on the pharmaceutical properties of medications. These properties cover different types of prescription information (e.g., medication name, frequency, duration, and dosage) and drug-related safety concerns (e.g., potential adverse drug reactions). An adverse drug reaction (ADR) describes an undesirable effect associated with the administration of a particular medication. The automatic extraction of ADRs serves two important purposes: (1) to compare ADRs extracted from different labels (i.e., across different manufacturers) that correspond to the same drug, and (2) to conduct pharmacovigilance by identifying new ADRs that are not currently described in the labels[1-2].

We propose a system that automatically extracts clinically relevant entities from drug labels, including: ADRs, ADR severities, drug classes, negated ADRs, animal species, and factors (any additional aspect of an ADR that is not covered by one of the other mentions) from drug labels. The first component of our system identifies non-negated ADRs. Then, for each non-negated ADR, our system provides MedDRA[3] Preferred Terms (PTs) and Lowest Level Terms (LLTs). When evaluated on the official test set from the 2017 TAC ADR shared tasks, our system achieved a micro-average F1-measure of 0.702 on the extraction of ADRs and related mentions (Task 1, exact match). The relation extraction module (Task 2) links together ADRs (from Task 1) with related mentions (i.e., Negated, Hypothetical, and Effect). This module achieved an F1 score of 0.310. Finally, the identification of positive ADRs (Task 3) and their normalization (Task 4) achieved F1 scores of 0.703 and 0.780 respectively.

2. Data

The dataset provided by the organizers of 2017 TAC ADR shared tasks contains 101 annotated drug labels and 2,208 unannotated drug labels (See **Table 1** for the per-category statistics).

Table 1 - Number of entities per category in the training and test data

Dataset	Record	ADR	Animal	Drug Class	Factor	Negation	Severity
annotated	101	12,323	44	244	600	87	760
unannotated	2,208	N/A	N/A	N/A	N/A	N/A	N/A

In addition to the 2017 TAC ADR dataset, we incorporated a manually curated database of side effects, called VigiBase, from VigiAccess.org[4]. VigiBase contains 18,310 ADRs commonly found in the drug prescriptions. We also used the Unified Medical Language System (UMLS) metathesaurus[5].

3. Methods

We approached the extraction of ADRs (and other clinically relevant entities) as a named entity recognition (NER) task. We trained conditional random fields (CRFs) to mark named entities that we were interested in detecting. We found the relationships between these entities through regularized regressions and filtered out negated or hypothetical ADRs using rules. We then normalized any ADRs extracted by our system to MedDRA PTs and LLTs using hand-written rules.

3.1 Pre-processing

The first step in our NER pipeline was to pre-process the given dataset. To do so, we performed sentence breaking on the drug labels by splitting on periods, and tokenized these sentences by splitting on whitespace. We did not split periods found in acronyms, lists, and numbers. We lowercased all tokens and used the Natural Language Toolkit (NLTK)[6] to assign part-of-speech (POS) tags. We also replaced numbers (in both numerical and literal forms) with placeholders (e.g., *grade 4 proteinuria → grade D proteinuria*, *1st degree atrioventricular block → D degree atrioventricular block*). Additionally, each document has been parsed using Stanford CoreNLP[18] with the aim of extracting lemmas, POS tags, constituency, and dependency parsing trees.

3.2 ADRs extraction

The goal of task 1 was to identify ADRs and ADR-related entities. Most high performance clinical NER models train on large feature sets[7-9]. In our experiments, we first attempted to use a smaller set of features, which contained only normalized tokens, POS tags, and real-valued word vectors. The efficacy of real-valued word embeddings in clinical NER[10-11] motivated our experiments with the real-valued-word vectors. We trained two such vectors on two different datasets using GloVe[12]. First, we trained 100-dimensional real-valued word vectors on the MIMIC III dataset[1]. Second, we trained 100-dimensional real-valued word vectors on the dataset provided for the 2017 TAC ADR Shared Tasks, which includes 2,309 drug labels. Note, if a token was not included in the pre-trained vector set, we assigned the vector of token <*unknown*> instead.

We trained two CRFs, one for each of the trained real-valued word vectors, on the annotated drug labels. In both models, the window size for tokens and POS tags is ± 2, which has demonstrated to be the optimal size[10], through 5-fold cross-validation.

We also trained a third CRFs-based model, with a much larger feature space, broader window size and more complex topology graph. This model included morphological, lexical and

[1] MIMIC III is a large critical care database that contains approximately 2 million clinical notes for approximately 46,000 patients[13].

syntactic features commonly used in a NER task[14]. The lexical features used gazetteers extracted from VigiBase. Each gazetteer corresponds to a specific disorder/condition type (e.g., *blood, cardiac, familial, ear-related*).

3.3 Relation Extraction

Task 2 aimed to automatically extract the relations between the ADRs and the related information identified in Task 1. The relations provided by the annotators involved two entities, where the first one was an ADR and the second was a mention of related information (severity, factor, drug class, negation, and animal). None of the documents in the training data included cross-sentence relations.

We characterized each pair of entities with features based on lexicon, distances, constituency and dependency trees. We included information such as the direction of the relation, the juxtaposition of the prepositions between the entities, the distance of the headwords measured in tokens and edges on the constituency and dependency tree. We also included information related to the entities involved in the relation: lemma, POS, textual form, type, and common ancestor (for multi-word entities). We trained a regularized linear regression model on the entire training data.

3.4 Positive ADRs Identification

The purpose of Task 3 was to extract non-negated ADRs from drug labels. To accomplish this task, we developed a rule-based filtering component built on top of the output produced by the first two tasks. Specifically, we filtered out ADRs having negated or hypothetical relationships to drug classes. We also filtered out ADRs that are related to animal species rather than humans.

3.5 Normalization of positive ADRs

The purpose of Task 4 was to normalize non-negated ADR entities extracted in previous tasks. To do so, we used a rule-based approach that exploited the UMLS Metathesaurus[5]. We decided to forgo training a machine learning model because of the limited amount of training data. However, the rule-based approach demonstrated to be effective due to the robustness of existing tools[15].

We used MetaMap and Sub-term mapping tools (STMT) as normalizer. MetaMap is a widely adopted tool that maps medical text to UMLS Metathesaurus Concept Unit Identifiers (CUIs)[16]. MetaMap provides various processing options which affect the mapping result. The STMT is a generic toolset designed to find all sub-terms in a specific corpus and map synonymic variations of the corpus to UMLS concepts[17]. Thus, we experimented with MetaMap and STMT by testing different configurations to achieve optimal normalization performance on the training set. We tested MetaMap and STMT using the 2016AA edition of UMLS knowledge source, which

aligned with the MedDRA v18.1 requirements for Task 4. Ultimately, we concluded that, on the training set, the NLM strict model yielded optimal performance when we processed terms by ignoring word order. For ADRs not matched to a CUI, we applied STMT. This tool maps a term to a relevant CUI by substituting a term in the given input to a synonymic term (i.e., if the term exists in the corpus of synonyms). For example, if "*Fetal Harm*" is given as a input, STMT finds the relevant CUI "*Foetal Damage*" by substituting the term "*harm*" to the synonymic term "*damage*". There is a tradeoff between precision and recall that needs to be optimized with the introduction of STMT. In our case, we decided to adopt STMT because it increased system recall, which improved the overall performance of our system on the training set.

Medical abbreviations are used frequently in the corpus of drug labels. These abbreviations introduce an ambiguity problem as multiple concepts can be represented by the same abbreviation. For example, the abbreviation "*SJS*" can be expanded to both "*Schwartz-Jampel Syndrome*" and "*Steven-Johnson Syndrome*". We observed that some ADRs are not normalized correctly because of improper abbreviation expansion. Thus, we implemented an abbreviation expander that uses context to determine appropriate expansion of the abbreviation. Specifically, we collected terms in the drug labels with more than two consecutive capital letters in parentheses to identify abbreviations. For each abbreviation identified, we collected several terms antecedent to the abbreviation to use as a potential expansion. We then mapped the expansion to its abbreviation, which we compiled into a dictionary. The dictionary was used only for ADRs extracted from the same drug label. After substituting abbreviations with expansions, we again executed our pipeline of MetaMap and STMT.

After using the UMLS Metathesaurus to find a MedDRA PT that corresponded to a CUI, we needed to determine if the MedDRA PT was the most relevant LLT. In the case that the MedDRA PT identified by the UMLS Metathesaurus was not the most relevant LLT, we needed to select which among multiple candidates was most relevant. We experimented with several methods, including exact text matching (with and without ignoring word order), and inexact matching. We found that exact text matching without ignoring word order demonstrated the highest performance on the training set.

4. Results

We evaluated the performance of our system using 5-fold cross-validation on the training set. We then trained our machine learning models on the entire training set and evaluated on the test set. The test set is a subset of 99 drug labels from 2,208 unannotated drug labels. The test set was provided by TAC organizers after our system submission without gold standard annotation. **Table 2** shows our system performance on Task 1 on both training and test sets. **Table 3** show our system performance on Task 2 using the test set. For task 1, the model using word embeddings trained on TAC drug labels outperformed the model using word embeddings trained

on the MIMIC III dataset. Thus, we used the model trained on TAC word embeddings to extract ADRs and other clinically relevant entities for the unannotated TAC challenge dataset.

Table 2 - F1-measure (exact match) of Task 1 on the training and test sets

Dataset	Vectors	ADR	Animal	Drug	Factor	Negation	Severity	Micro-Avg
Training	MIMIC III	0.756	0.798	0.155	0.523	0.258	0.587	0.730
Training	TAC	0.762	0.786	0.143	0.532	0.309	0.592	0.735
Test	TAC	N/A	N/A	N/A	N/A	N/A	N/A	0.702

Table 3 - F1-measure of Task 2 on the test set

Task	Precision	Recall	F1-measure
Relation Extraction	0.505	0.224	0.310

Table 4 shows the system performance for Task 3 on both training and test sets. As mentioned, the performance of Task 3 is entirely dependent on the performance of Task 1 and Task 2 because it relies on the output of those two tasks.

Table 4 - F1-measure of Task 3 on the training and test sets

Dataset	Precision	Recall	F1-measure
Training	1.000	1.000	1.000
Test	0.732	0.689	0.703

The evaluation for the Task 4 is presented in **Table 5**. Without incorporating other tools, MetaMap produce high precision (0.907) but low recall (0.796) on the training set.

Table 5 - F1-measure of Task 4 on the training and test sets

Dataset	Precision	Recall	F1-measure
Training	0.900	0.809	0.852
Test	0.853	0.728	0.780

Although the introduction of STMT decreases system precision, it improves recall to such a degree that overall performance improved. Furthermore, by expanding abbreviations through the use of our customized dictionary, we further improved the recall of our system. Using MetaMap, STMT and abbreviation expansion produces the highest performance of 0.780 on the test set. The system performance evaluated on the test set is lower than that on the training set because the results from previous tracks were imperfect.

5. Discussion

Our system cannot extract entities containing multiple or overlapped phrases. For example, "*increased alanine transaminase (ALT)*" contains two ADRs as suggested by manual annotation (i.e., "*increased alanine transaminase*" and "*increased ALT*"). Our system cannot extract these entities when "*increased*" is shared by both. Also, in "*exacerbation of pre-existing diabetes mellitus*", the extraction is problematic because "*exacerbation diabetes mellitus*" is separated into two sequences. Given imperfect output from Task 1, low performance for the relation extraction task (i.e., Task 2) partially resulted from the errors cascading from Task 1 (i.e., the extraction of ADRs). For Task 3, as seen on Table 4, our system provides perfect output given perfect input. Therefore, we conclude that the errors on the test set are due to the errors in the outputs to Task 1 and Task 2.

The evaluation of Task 4 is based on how many [PT , LLT] pairs the system accurately detects. Therefore, the system performance is highly dependent on selection of the most relevant LLTs among multiple candidates (i.e., even after the system identifies a relevant PT). However, in some cases, we did not properly select a relevant LLT. For example, when "*skin dry*" was input to the system, our system did identify the correct corresponding CUI (i.e., *C0151908*). The concept is mapped with one PT ("*10013786 skin dry*") and one LLT ("*10040835 dry skin*") in UMLS Metathesaurus. Because there is no difference between "*skin dry*" and "*dry skin*", ("*10013786", "10040835*") pair should be included in the training set, but ("*10013786", None)* was suggested as a correct normalization.

6. Conclusion

We propose an end-to-end system for the automatic extraction of ADRs and other clinically relevant entities, their relations, negation filtering, and mapping to MedDRA PTs and LLTs. The system used CRF-based models to extract ADRs and related mentions, linear regression to extract relations, and MetaMap with STMT to query and map extracted ADRs into normalized forms. When evaluated on the official test set from the 2017 TAC ADR shared tasks, our system achieved a micro-average F1-measure of 0.702 on the extraction of ADRs and related mentions (exact match). The relation extraction module achieved an F1 score of 0.310. Finally, the

identification of positive ADRs (Task 3) and their normalization (Task 4) achieved an F1 score of 0.703 and 0.780, respectively.

References

[1] Food and Drug Administration. Guidance for Industry-Adverse Reactions Section of Labeling for Human Prescription Drug and Biological Products—Content and Format. Rockville, MD: US Department of Health and Human Services. 2006.

[2] Guidance D. Guidance for Industry. Center for Drug Evaluation and Research (CDER). 2013 Feb;37:38.

[3] MedDRA MS. Introductory Guide MedDRA Version 17.1. Chantilly, VA: MedDRA Maintenance and Support Services Organization. 2014.

[4] VigiAccess. (n.d.). Retrieved October 26, 2017, from http://www.vigiaccess.org/

[5] UMLS - Metathesaurus. (n.d.). Retrieved October 26, 2017, from https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/

[5] Loper E, Bird S. NLTK: The natural language toolkit. InProceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1 2002 Jul 7 (pp. 63-70). Association for Computational Linguistics.

[7] Li, D., Kipper-Schuler, K. and Savova, G., 2008, June. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In Proceedings of the workshop on current trends in biomedical natural language processing (pp. 94-95). Association for Computational Linguistics.

[8] Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the American Medical Informatics Association. 2011 Apr 20;18(5):601-6.

[9] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011 Jun 16;18(5):552-6.

[10] Tao C, Filannino M, Uzuner Ö. Prescription extraction using CRFs and word embeddings. Journal of Biomedical Informatics. 2017 Aug 1;72:60-6.

[11] Wu Y, Xu J, Jiang M, Zhang Y, Xu H. A study of neural word embeddings for named entity recognition in clinical text. In AMIA Annual Symposium Proceedings 2015 (Vol. 2015, p. 1326). American Medical Informatics Association.

[12] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 (pp. 1532-1543).

[13] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3.

[14] Filannino M, Nenadic G. Temporal expression extraction with extensive feature type selection and a posteriori label adjustment. Data & Knowledge Engineering. 2015 Nov 30;100:19-33.

[15] Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. Journal of the American Medical Informatics Association. 2012 Oct 6;20(5):876-81.

[16] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. InProceedings of the AMIA Symposium 2001 (p. 17). American Medical Informatics Association.

[17] Lu CJ, Browne AC. Development of Sub-Term Mapping Tools (STMT). In AMIA 2012.

[18] Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The stanford corenlp natural language processing toolkit. InACL (System Demonstrations) 2014 Jun 23 (pp. 55-60).