Extracting Adverse Drug Reactions from Forum Posts and Linking them to Drugs

Andrew Yates
Information Retrieval Lab
Department of Computer Science
Georgetown University
andrew@ir.cs.georgetown.edu

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

Ophir Frieder
Information Retrieval Lab
Department of Computer Science
Georgetown University
ophir@ir.cs.georgetown.edu

ABSTRACT

Interest in medical data mining is growing rapidly as more health-related data becomes available online. We propose methods for extracting Adverse Drug Reactions (ADRs) from forum posts and linking extracted ADRs to the drugs that users claim are responsible for them. We evaluate our methodology using a corpus of annotated forum posts. We find that our ADR extraction method outperforms a strong baseline in terms of precision at the expense of a similar decrease in recall. When used in conjunction with a strong baseline, our method is able to increase recall by 7% without harming precision.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation

Keywords

Medical data mining, extracting adverse drug reaction, health-related social media

1. INTRODUCTION

According to a recent survey, 35% of US adults have attempted to use the Internet to diagnose a medical condition in the past year[1]. As more health-related data appear online, interest in medical data mining grows rapidly. Recent efforts include those that mine symptoms and conditions from query logs [13], mine adverse drug reactions from drug reviews [10, 15], health-related social networks [9], and forum posts [2], and mine the existence of particular [1, 4, 5] or any [11] medical outbreaks.

We focus on extracting Adverse Drug Reactions (ADRs) from forum posts and linking the extracted ADRs to the drug that the user claims, implicitly or explicitly, is responsible for them. Forum posts present unique challenges in that they are relatively long (130 terms on average, in our dataset) and the relationships between drugs and ADRs are not available as structured data. In contrast, drug comments posted to some other forms of social media, such as drug review sites, are explicitly related to a specific drug (i.e., the review or comment is posted on a drug's page) and are relatively succinct (46 terms on average, in the dataset used in [15]).

HSD 2013, August 1, 2013, Dublin, Ireland.

Specifically, we propose 1) a novel method for extracting ADRs from social media using linguistic dependency relations and a conditional random field (CRF) [8], and 2) a novel method for linking ADRs to the drugs that posting users identify as their cause.

Our contributions are

- A novel method for extracting ADRs from social media
- A novel method for linking ADR mentions to drugs
- A publicly available annotated dataset[2] indicating the ADRs present in forum posts and the drugs that users identified as being responsible for the ADRs.

2. RELATED WORK

Several previous efforts focused on extracting adverse drug reactions (ADR) from social media.

Leaman et al. [9] matched terms in a bag-of-words sliding window against known ADRs after correcting for spelling mistakes. Similarly, Benton et al. [2] found ADRs occurring in sets of terms that were more likely to occur together within a bag-of-words sliding window than they were to occur separately. We compare our approach to a bag-of-words sliding window approach.

Li [10] used statistical methods to find terms that were only present in one of two mutually exclusive classes of drug. This method requires that two mutually exclusive drug classes be compared, which requires domain knowledge and is not possible when two such classes do not exist.

Yates & Goharian [15] proposed ADRTrace, a system that found ADRs that exactly matched terms in a list or matched a pattern mined from drug reviews. We compare our approach to ADRTrace.

3. METHODOLOGY

We describe our method for extracting adverse drug reactions (ADRs) in Section 3.1. In Section 3.2 we describe our method for associating extracted ADRs with the drug that the user claims is responsible for the ADR.

[1] http://www.pewinternet.org/Reports/2013/Health-online.aspx

[2] http://ir.cs.georgetown.edu/data/adr_forum_annotations

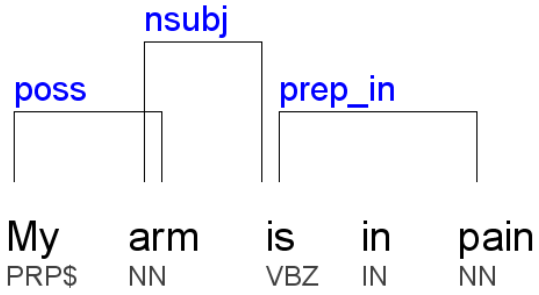

Figure 1. Example dependency relations

3.1 Extracting ADRs

Rather than attempting to match all terms in a sliding window against known ADRs as some previous efforts have done, we focus on using dependency relations as a principled way to choose terms to match against known ADRs. We use the Stanford Parser [7] to identify dependency relations, which consist of a relation type (e.g., nominal subject), a head term (i.e., the term which determines the type of phrase), and a modifier term (i.e., a term which modifies the head term). For example, Figure 1 shows the collapsed dependency relations from the sentence "My arm is in pain." A sliding window of at least 4 terms would discover the "arm pain" ADR in this sentence, but it is also possible to discover that ADR by noting that "arm," "is," and "pain" are connected by relations and checking for ADRs composed of a subset of those words.

We find ADRs using dependency relations in two ways. First, we find ADRs by combining every pair of terms that appears in a dependency relation and determine whether the term pair matches a known ADR. If it does the ADR is extracted.

Second, we learn which dependency relation paths can be followed to generate candidate ADRs. To do so, we construct a graph from each forum post. Each vertex corresponds to a term; edges correspond to dependency relations between terms. Figure 2 shows a subset of one such graph. The post contains the "joint pain," "ankle pain," and "fatigue" ADRs, among others. The edges are labeled with the dependency relation types connecting the term vertices (e.g., "amod" and "dep"). It is our method's job to determine when these edges can be followed; we use the relation types (e.g., "amod") later as one of the features used to determine that. This example is kept small so that it may be easily visualized; in actual use ADRs may consist of several terms that are each several hops away from each other.

Each post is then split into individual sentences using the Punkt sentence tokenizer [6]. Each sentence is treated as bag-of-words to find potential ADRs that may exist in the sentence (e.g., "carpal tunnel syndrome" would be found in "syndrome x carpal y tunnel z"). For each of these potential ADRs, we find the shortest path in the graph between each sequential pair of terms (e.g., we find the shortest path between "carpal" and tunnel" and the shortest path between "tunnel and syndrome").

Each sequence of shortest paths in a potential ADR is then turned into binary features for use with a Conditional Random Field (CRF) [8], which are often applied to classification tasks involving natural language, such as named entity recognition. Note that there may not be a single hop path between each term in a potential ADR; thus, term vertices may appear in the path between ADR terms that are not included in the extracted ADR.

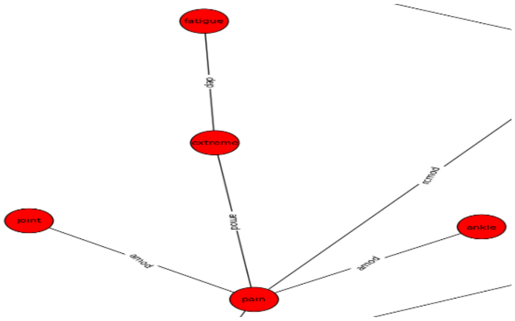

Figure 2. Dependency relation graph

The features used by the CRF are:

- Dependency relations present in the path
- Terms present in the path (no distinction is made between head terms and modifier terms)
- Term appearance anywhere in the MedSyn thesaurus [15]
- Head term appearance anywhere in the MedSyn thesaurus
- Modifier term appearance anywhere in the MedSyn thesaurus

These features capture the terms and dependency relations present in the path, and whether each term could be part of an ADR (i.e., whether it exists in MedSyn). This allows the CRF to determine if the path should be followed. When training the CRF, each set of features corresponding to an ADR that is known to exist in a post is given the "FOLLOW" label. Each set of features that corresponds to an ADR that does not exist but could (i.e., a bag-of-words sliding window method with a large window would extract the ADR but the ADR is actually not expressed in the post) is given a "DON'T_FOLLOW" label.

To find ADRs using this method, the CRF is given a set of features for each ADR whose terms exist in the post (i.e., each ADR that would be found using a sliding window approach when the window size is equal to the post's length). The ADR is extracted if the CRF predicts the "FOLLOW" label.

3.2 Linking Drugs to ADRs

While it is impossible to determine whether a relationship expressed between a drug and an ADR is true, namely true causation, it is still helpful to detect when such a relationship is expressed. For example, a relationship between the drug "Tamoxifen" and the ADR "hot flashes" is expressed in the sentence "I've been having hot flashes since I started taking Tamoxifen." We detect such relationships by using a CRF to label drug mentions as "DRUG-<drug name>" and ADRs linked to the drug as "<drug name>-ADR." The CRF is run on terms that are each given the following binary features:

- Term itself
- Part-of-speech tag for the three terms before the current term and after the current term (e.g., "VRB@-3"). The tags were found using the Stanford Log-linear Part-of-Speech Tagger [12]. Three terms were used because this performed better than using two terms; increasing the window size to four terms did not improve performance.
- The part-of-speech tag for the current term

- The appearance of the current term as a term anywhere in the thesaurus
- The current term matches the name of a known drug
- The types of the dependency relations that the term appears in as either a head or modifier term anywhere in the post (e.g., "dobj, nsubj")

The CRF is used to label ADR's caused by a drug as "<drug name>-ADR." Any ADR with a term bearing the "<drug name>-ADR" label is linked to the drug <drug name> by our system.

4. EVALUATION

We describe our dataset in Section 4.1. We use this dataset to evaluate our ADR extraction performance in Section 4.2 and to evaluate our drug linking performance in Section 4.3.

4.1 Dataset

Evaluating our algorithms requires a corpus from which to extract ADRs and the drugs associated with them, a domain-specific thesaurus listing terms and phrases that are equivalent (i.e., refer to the same ADR), and annotations indicating the ADRs expressed in the corpus and the drugs associated with these ADRs.

We used a corpus consisting of 400,000 posts crawled from the Breastcancer.org[3] and FORCE[4] forums. The posts were primarily chosen from sub-forums related to the discussion of ADRs caused by breast cancer drugs. Information on obtaining the corpus is available in [14]. MedSyn [15], a list of synonyms in the medical ADR domain that includes both expert (e.g., "arthralgia") and non-expert (e.g., "joint pain") terms, was used as our thesaurus. MedSyn is derived from a subset of the Unified Medical Language System Metathesaurus (UMLS) [3]; a description of how MedSyn was constructed is available in [15].

The corpus posts were annotated. Non-medically trained annotators were asked to read posts, annotate the ADRs present in the posts, and, if mentioned, indicate the drug that the user associated with each ADR. Annotators were instructed to only annotate first-hand accounts of an ADR and allowed to skip ADRs that were related to a medical procedure (e.g., surgery or chemotherapy). Each post was annotated by three separate annotators; posts and annotations that did not meet this criterion were discarded. In total, the annotators annotated approximately 600 posts with a total of 2,100 annotations. MedSyn was used to treat different terms or phrases that expressed the same ADR as equivalent. Fleiss' Kappa was calculated to be 0.37, indicating fair inter-rater reliability. To use these annotations as ground truth, we discarded any ADRs that were found by only one annotator. When evaluating our ADR extraction performance, we only included the 1,700 annotations that were related to one of the following breast cancer drugs: Arimidex, Aromasin, Femara, Nolvadex, and Taxotere. The annotations and URLs of the forum posts are available on our website[5].

We used the annotations to evaluate our methods rather than using the ADRs listed on drug labels. While we expect that some of the ADRs listed on drug labels are also expressed in the forum posts, other listed ADRs may be infrequent and should not be expected to be found. Annotations are used to perform a more direct evaluation by comparing the ADRs and drug relationships we extract to the annotated ADRs and relationships. Furthermore, the annotations may be used to train and evaluate supervised methods when coupled with the forum posts they correspond to.

4.2 ADR Extraction

We used our corpus and annotations to evaluate ADR extraction performance. Five-fold cross-validation was used.

The results, the baselines, and the *DepADR* system described in this paper are shown in Table 1. The percentages in parentheses indicate a system's performance relative to *ADRTrace*'s. An asterisk (*****) indicates a statistically significant change in performance at $p < 0.05$. *ADRTrace+DepADR* indicates that every ADR extracted by either the *ADRTrace* or *DepADR* system is returned. *Window* is a bag-of-words sliding window system with sliding windows of size 25; we chose a large window to establish an upper-bound on recall.

Window achieves the highest recall, as would be expected. Using *DepADR* in conjunction with *ADRTrace* yields a 7% increase in recall without harming precision, supporting our hypothesis that dependency relations can be used to extract additional ADRs without returning many more false positives. When used by itself, *DepADR* achieves a 56% improvement in precision at the expense of a 48% reduction in recall. This is unsurprising given *DepADR*'s focus on carefully choosing which terms may compose ADRs. These results suggest that *DepADR* can be paired with an existing system to improve performance when recall is important or used by itself in scenarios where a higher precision is desired.

Table 1. ADR extraction results

	Precision	*Recall*
ADRTrace	0.39	0.61
ADRTrace+DepADR	0.39	0.65 (+7%)*****
DepADR	0.61 (+56%)*****	0.32 (-48%)*****
Window	0.32 (-18%)*****	0.74 (+21%)*****

4.3 Drug Linking

We used our corpus and annotations to evaluate our method for linking drugs to ADRs. Five-fold cross-validation was used. We compare our performance to a baseline that returns all (drug, ADR) pairs that exist in a post.

The results are shown in Table 2. The percentages in parentheses indicate a system's performance relative to *Baseline*'s. An asterisk (*****) indicates a statistically significant change in performance at $p < 0.05$. The baseline outperforms our system in terms of recall, which is to be expected given that the baseline returns every possible link. Our system performs 17% better in terms of precision. Coupled with our relatively low recall, these results suggest that our CRF approach is promising but could be improved.

Table 2. ADR-Drug linking results

	Precision	*Recall*
DepADR Linking	0.63 (+17%)*****	0.36 (-64%)*****
Baseline	0.54	1.0

[3] http://community.breastcancer.org/

[4] http://www.facingourrisk.org/messageboard/index.php

[5] http://ir.cs.georgetown.edu/data/adr_forum_annotations

5. CONCLUSIONS

We proposed novel methods for extracting ADRs from social media and linking each ADR to the drug that the user identified as being responsible for the ADR. We use supervised methods for both tasks that are trained on our annotated dataset. Our ADR extraction method makes extensive use of dependency relations to precisely choose potential terms to match against ADRs. Our results show that our ADR extraction method statistically significantly outperforms two previously proposed methods, while our drug linking method outperforms a simple baseline.

This work is clearly preliminary; future work will refine the approach described and combine various approaches to improve precision without significantly hampering recall.

6. REFERENCES

[1] Aramaki, E. et al. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. *Conference on Empirical Methods in Natural Language Processing (EMNLP'11)* (Jul. 2011), 1568–1576.

[2] Benton, A. et al. 2011. Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *Journal of biomedical informatics*. 44, 6 (Dec. 2011), 989–96.

[3] Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*. 32, Database issue (Jan. 2004), D267–70.

[4] Corley, C. et al. 2009. Monitoring Influenza Trends through Mining Social Media. *International Conference on Bioinformatics Computational Biology (ICBCB'09)* (2009).

[5] Jamison-Powell, S. et al. 2012. "I can't get no sleep": discussing #insomnia on twitter. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (New York, New York, USA, May. 2012), 1501.

[6] Kiss, T. and Strunk, J. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*. 32, 4 (Dec. 2006), 485–525.

[7] Klein, D. and Manning, C.D. 2003. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03* (Morristown, NJ, USA, Jul. 2003), 423–430.

[8] Lafferty, J.D. et al. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Eighteenth International Conference on Machine Learning (ICML'01)* (Jun. 2001), 282–289.

[9] Leaman, R. et al. 2010. Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. *2010 Workshop on Biomedical Natural Language Processing (BioNLP'10)* (2010), 117–125.

[10] Li, Y.A. 2011. *Medical Data Mining : Improving Information Accessibility using Online Patient Drug Reviews*. MIT.

[11] Parker, J. et al. 2013. A Framework for Detecting Public Health Trends with Twitter. *IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM'13)* (2013).

[12] Toutanova, K. et al. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03* (Morristown, NJ, USA, May. 2003), 173–180.

[13] White, R.W. and Horvitz, E. 2012. Studies of the onset and persistence of medical concerns in search logs. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12* (New York, New York, USA, 2012), 265.

[14] Yates, A. et al. 2013. Graded relevance ranking for synonym discovery. *22nd international conference on World Wide Web companion (WWW '13 Companion)* (May. 2013), 139–140.

[15] Yates, A. and Goharian, N. 2013. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. *Proceedings of the 35th European conference on Advances in Information Retrieval (ECIR'13)* (2013).