ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites

Andrew Yates and Nazli Goharian

Department of Computer Science
Georgetown University

{andrew,nazli}@ir.cs.georgetown.edu

Abstract. We automatically extract adverse drug reactions (ADRs) from consumer reviews provided on various drug social media sites to identify adverse reactions not reported by the United States Food and Drug Administration (FDA) but touted by consumers. We utilize various lexicons, identify patterns, and generate a synonym set that includes variations of medical terms. We identify "expected" and "unexpected" ADRs. Background (drug) language is utilized to evaluate the strength of the detected unexpected ADRs. Evaluation results for our synonym set and ADR extraction are promising.

1 Introduction & Prior Work

According to information from a Pew Internet Project Survey conducted in 2010[1], 80% percent of Web users look for health-related information online. We are interested in harvesting this information to identify both expected (i.e., known) and unexpected adverse drug reactions (ADRs). It is important to note that we do not claim the unexpected ADRs are necessarily caused by the medications, but are conditions that should be tracked medically by a domain expert.

Our effort is motivated by [2] to mine social media for pharmacovigilance. Two prior efforts [4,5] mine social media to extract ADRs. Unlike our effort, the authors in [4] mined ADRs from drug discussion forums using a sliding window, suffering from inexact matching. More so, we also differ by using "background drugs" to differentiate detected unexpected ADRs that might be caused by the condition being treated rather than the drug being used.

[1] http://pewinternet.org/Commentary/2011/November/Pew-Internet-Health.aspx

The authors in [5] extract ADRs for drugs from user reviews by using statistical methods to compare the terms present in two mutually exclusive classes of drug. We differ by requiring less domain knowledge and accounting for ADR synonyms.

The lack of an annotated benchmark dataset complicates our approach. Thus, we created a publicly available annotated breast cancer drug review dataset (available at http://ir.cs.georgetown.edu/data/adr). We also generated a synonym set to address the language gap between users and medical professionals (e.g., "joint pain" and "arthralgia" refer to the same ADR).

Our contributions are: (1) creating an annotated breast cancer drug review dataset; (2) generating a comprehensive ADR synonym set focused on breast cancer (MedSyn); and (3) extracting expected and unexpected ADRs from drug reviews (ADRTrace).

2 Dataset

We crawled 2500 user reviews for five commonly used breast cancer drugs: Anastrozole, Exemestane, Letrozole, Raloxifene, and Tamoxifen. These reviews were collected from three drug review social media sites, namely, askapatient.com, drugs.com, and drugratingz.com. We annotated 10% of the reviews for adverse drug reactions (ADRs); 50% of this was used to generate a synonym set (training data) and the other 50% was used to evaluate the quality of the ADR extraction (testing data).

We used SIDER 2 [3], a resource containing known ADRs, to determine whether ADRs were expected or unexpected. Unified Medical Language System (UMLS) is a set of medical lexicons and a semantic network that expresses the relationships among terms. Our synonym set, which we refer to as "MedSyn," was built using a subset of UMLS that contained 5 relevant medical lexicons with approximately 208,000 terms (both consumer and medical).

3 Architecture

Our ADRTrace system consists of two main components: an ADR synonym set and a mining engine. Our synonym set generator creates equivalent classes for ADR terms. The mining engine extracts expected & unexpected ADRs from reviews.

3.1 Generating the MedSyn Synonym Set

We based the MedSyn synonym set on a subset of UMLS. We used SIDER 2 to determine which UMLS semantic types could be ADRs (e.g., "disease or syndrome") and added every term with one of these semantic types to MedSyn. We then used

UMLS' semantic network to determine which terms in MedSyn were siblings and treated them as MedSyn synonyms. Finally, we used the training set to identify common ADR terms that were not already included and added them. MedSyn consists of 5,000 synonym groups and 30,000 synonyms, with an average of 5.5 synonyms per group ($\sigma = 4.8$).

3.2 Extracting ADRs

We extracted ADRs from online reviews by identifying review terms and phrases that appeared in our MedSyn synonym set. To improve the extraction, we utilized 7 patterns, which were handcrafted with the help of pattern extraction [6], to identify ADRs that did not appear in MedSyn. An example of such a pattern is "<X> in my <area> and <area>", which finds the "hip pain" and "joint pain" ADRs in the text "pain in my hips and joints." For brevity, we omit the other patterns. The ADRs extracted using a pattern were required to appear in MedSyn.

The SIDER 2 database was used to determine whether the extracted ADRs were expected or unexpected. We used two common cancer drugs, Premarin and Provera, as "background drugs" to separate reactions that were potentially related to cancer from ADRs that were reactions to one of the breast cancer drugs. We calculate the *support* and *strength* of each (unexpected ADR, drug) pair as shown below. A *strength* greater than zero suggests that the ADR is not caused by the underlying condition.

$$Support_{ADR,reviews} = \frac{|\{ADR \mid ADR \in reviews\}|}{|reviews|}$$

$$Strength_{ADR,reviews} = Support_{ADR,reviews} - \frac{\sum_{D \ni background} Support_{ADR,D}}{|background|}$$

4 Results & Evaluation

ADR extraction was evaluated using our annotated data set. The training set was used to develop the ADRTrace mining engine, while the testing set was used only to evaluate the engine's performance. We evaluated the portion of our MedSyn synonym set that was used to extract ADRs by manually inspecting the unexpected ADRs found and determining whether they were synonyms for expected ADRs.

Table 1 depicts precision, recall and F1 for the ADR extraction and synonym set evaluation. We achieve a high ADR extraction recall and relatively good precision. MedSyn achieves a high precision, indicating that the majority of extracted ADRs identified as unexpected ADRs are not synonyms of expected ADRs. Table 2 depicts a selection of unexpected ADRs identified using the entire data set. Background support indicates the ADR's support in the 1,300 background drug reviews . "Weight loss" occurs with equal support in both the breast cancer drugs and background drugs, giving it a strength of 0 and suggesting that ADR may be related to other cancer symptom/treatment and not to these drugs.

Table 1. Evaluation Results

Precision	*Recall*	*F1*	*Data*
0.69	0.89	0.78	ADR Extraction (Testing Set)
0.72	0.90	0.80	ADR Extraction (Training Set)
0.86	N/A	N/A	MedSyn (Synonym Set)

Table 2. Support, Background Support, and Strength for Detected Unexpected ADRs

Drug	*Supp. (#)*	*Bg. Supp. (#)*	*Strength*	*ADR*
An.	0.04 (43)	-	0.04	Osteopenia
An., Ex., Ra., Ta.	0.03 (63)	0.01 (7)	0.02	Memory loss
Le. & Ra.	0.03 (20)	-	0.03	Skin dryness
Ex. & Ra.	0.02 (6)	0.02 (25)	0	Weight loss

5 Conclusions

We presented a promising methodology for extracting expected and unexpected ADRs from social media reviews by using our MedSyn synonym set and ADRTrace mining engine.

6 References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. J. Nucleic Acids Research. 32, 267-270 (2004)

2. Federoff, H., Frieder O.: System and Method for Performing Pharmacovigilance. Patent Pending. (2012)

3. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., Bork, P.: A side effect resource to capture phenotypic effects of drugs. J. Molecular Systems Biology 6:343 (2010)

4. Leaman, R., et al..: Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In: Proc. of BioNLP '10. (2010)

5. Li, A. Y.: Medical Data Mining: Improving Information Accessibility using Online Patient Drug Reviews. Thesis. Massachusetts Institute of Technology. (2011)

6. Pantel P., Pennacchiotti, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: Proc. of ACL. (2006)