Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Learning the Relationships between Drug, Symptom, and Medical Condition Mentions in Social Media

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

We consider the general problem of learning relationships between drugs, symptoms, and medical conditions mentioned on Twitter, with the goal of estimating probability distributions to reduce the difficulties presented by social media's incomplete picture. If a user mentions taking a drug and experiencing several unexpected symptoms, for example, are the symptoms associated with that drug or is it more likely that the symptoms are associated with an unmentioned underlying condition? We describe a model for learning from and utilizing such relationships. We demonstrate that our approach identifies drugs that are similar based on their associated symptoms (or conditions), identifies conditions that are similar based on their associated symptoms, and can determine whether a symptom is caused by a medical condition or by a drug (i.e., a drug side effect).

Introduction

Social media data are subject to many biases, and sampled data sources such as Twitter's streaming API compound the problem. There is no guarantee that users will mention all details that are relevant to the mining task being performed, and even when users do provide complete information, the public Twitter API's sampling may prevent all of the tweets containing relevant information from being collected.

We reduce the difficulty of mining health-related data with incomplete information by modeling the relationships between drugs, symptoms, and medical conditions. If a user mentions suffering from a medical condition and experiencing a symptom, for example, the symptom may be caused by the medical condition or it may be caused by an unmentioned drug the user is taking. Our model addresses this problem by assessing the probability that a user who mentioned a condition and symptom is also associated with an unmentioned drug. Similarly, we demonstrate how our model estimates the similarity between drugs, symptoms, and conditions and determines whether symptom mentions are more likely to be associated with a condition mention or with a drug mention (i.e., drug side effects).

Many have considered the problem of modeling health-related latent topics in social media. Chen et al. (2014) use a temporal topic model to model users' flu infection statuses (i.e., healthy, exposed, or infected) over time based on the users' tweets. Paul and Dredze (2011) propose the Ailment Topic Aspect Model (ATAM+) to associate treatment, symptom, and general terms with latent ailment topics. While our goal of learning the associations between drug, symptom, and condition mentions is similar to modeling health-related topics, we differ in that we do not try to discover latent topics. We use concept extraction and medical thesauri to identify mentions rather than training a topic model to discover topics (i.e., term categories). We envision our model as augmenting health-related mining tasks such as discovering drug side effects or estimating the prevalence of a disease.

Methodology

We model relationships between symptoms, drugs, and medical condition mentions in tweets as a Bayesian network. A user's medical conditions determine what drugs the user takes. The user's conditions and drugs determine what symptoms the user experiences; a symptom may be caused by the condition or it may be a side effect of a drug the user is taking. We compute a joint probability distribution over symptoms, drugs, and conditions and use it to compute conditional probability distributions among them. We identify symptoms, drugs, and conditions in tweets using the CRF method described in (Yates, Goharian, and Frieder 2015).

Let S be a random variable over all symptoms, D be a random variable over all drugs, and C be a random variable over all conditions. Let $Count_{S,D,C;U}$ be the number of times S, D, and C are mentioned by the Twitter user U during a d-day window. We use nil values in cases where a d-day window contains only one or two types of random variable; that is, if a user mentions one or more symptoms or conditions during a window but does not mention any drugs, we set $D = \emptyset$ for that window. We set $d = 7$ in our experiments. The joint probability mass function is:

$$\Pr(S, D, C) = \frac{\sum_{U \in users} Count_{S,D,C;U}}{\sum_{S,D,C,U \in users} Count_{S,D,C;U}} \quad (1)$$

Conditional probabilities between any two of the random variables are then computed by marginalizing out the third variable. The conditional probability of extracting the symptom S given a drug D, for example, is:

$$\Pr(S|D) = \frac{\Pr(S,D)}{\Pr(D)} = \frac{\sum_C \Pr(S,D,C)}{\sum_{C,S} \Pr(S,D,C)} \quad (2)$$

Conditional probability distributions can be inspected to identify associations between symptoms, drugs, and conditions. The Kullback–Leibler divergence between distributions can be used to compare the similarity of random variables, such as comparing the similarity of symptoms associated with two drugs D_1 and D_2:

$$D_{\mathrm{KL}}(\Pr(S|D_1)||\Pr(S|D_2)) \quad (3)$$

Finally, we can identify symptoms that are either more highly associated with a condition than a drug (i.e., are symptoms of the condition) or more highly associated with a drug than a condition (i.e., are drug side effects) by taking the difference of the given drug's and condition's conditional probabilities:

$$\Pr(S|D_1) - \Pr(S|C_1) \quad (4)$$

Dataset

Our dataset consists of a thesauri containing symptom, drug, and condition terms and a Twitter corpus collected between November 2013 and November 2015. Rather than using Twitter's streaming 1% sample API, we used Twitter's `statuses/filter` API and queried for tweets geolocated within the United States and Canada to maximize our per-user coverage. Our Twitter corpus contains approximately 1.5 billion tweets written by 11 million users, of which about 18 million (1.2%) tweets contained a term from our final thesauri (i.e., were health-related tweets). Most health-related tweets mentioned at least one symptom (57%) or condition (37%), with only 7.6% of health-related tweets mentioning a drug. We use the SIDER (Kuhn et al. 2016) drug database to identify drug terms in our corpus. We use the MedSyn thesaurus (Yates and Goharian 2013), a thesaurus containing both lay person and expert terminology derived from the Unified Medical Language System (Bodenreider 2004), to identify health-related terms that may express symptom or condition concepts. Additionally, we manually verified every symptom, drug, and condition term that occurred at least 400 times in our corpus and removed ambiguous terms from our thesauri.

Experiments

Drug, symptom, and condition associations

To evaluate how well we learn associations between drugs, symptoms, and conditions, we compute the conditional probability (Eq. 2) of symptoms given NSAIDs (nonsteroidal anti-inflammatory drugs), which are a common class of over-the-counter painkilling drugs (i.e., $\Pr(S|D)$). The top ten symptoms associated with each drug are:

Advil headache (0.11), confused (0.08), sleepy (0.07), throw up (0.06), pass out (0.04), cough, cramps, feel sick, hangover & fever (0.03)

Aleve headache (0.10), sleepy (0.06), confused (0.06), gnecomastia (0.04), cramps (0.04), throw up (0.04), feel sick (0.03) , ache (0.03), fever (0.03) & cough (0.02)

Aspirin nasal polyps (0.08), polyps (0.08), tumor (0.05), swelling (0.05), swollen (0.04), headache (0.03), salivary glands (0.03), runny nose (0.03), fever & apoptosis (0.02)

Tylenol asthma (0.06), headache (0.06), fever (0.04), confused (0.04), hard of hearing, throw up, hearing loss, tumor, sleepy & migraine (0.03)

Symptoms that NSAIDs are commonly taken to relieve, such as headaches and fevers, are associated with every drug. Symptoms of underlying conditions that NSAIDs do not cause or treat also appear, however, such as cough, sleepy, and runny nose. This illustrates that while conditional probability can be used to find associations between drugs and symptoms, the association may be an indirect association that also involves an underlying condition. We demonstrate in a later section that Eq. 4 can be used to separate symptoms caused by an underlying condition from symptoms caused by a drug (i.e., drug side effects).

Table 1 shows the symptoms and drugs most strongly associated with four common conditions (i.e., those symptoms and drugs with the highest conditional probabilities given one of the conditions). Many of the symptoms are clearly symptoms of the condition: migraine, headache, sneeze, swollen, and cough are symptoms of allergies; tumors, polyps, and weight loss are symptoms of breast cancer, nausea (feeling sick), fever, and headache are flu symptoms, and shoulder pain is commonly associated with strokes. The relationships between conditions and drugs are less clear, with alcohol, cocaine, and testosterone commonly appearing. Allergies are correctly associated with allergy medications (i.e., zyrtec, prednisone, benadryl, and histamine), however, as well as a drug that alleviates an allergy symptoms (imitrex). Tamoxifen, one of the most common breast cancer drugs, ranks highly for the breast cancer condition. Tylenol, which ranks highly for the flu, is not associated with flu treatment but is used to alleviate some flu symptoms. The model's difficulty identifying drugs associated with breast cancer, the flu, and strokes may be caused by the fact that no drugs are commonly taken for these conditions; the flu is a common condition with no cure. Drugs are used in the treatment of breast cancer and strokes, but these are relatively rare conditions so their associated drugs are less likely to be mentioned on Twitter.

Similarities between drugs

To evaluate how well our model can be used to measure the similarity between two drugs, we compute the KL divergence (Eq. 3) between NSAIDs (nonsteroidal anti-inflammatory drugs). NSAIDs are commonly referred to both by their brand names and generic names, making them ideal for evaluating our drug-similarity metric. The KL divergences between pairs of NSAIDs are shown in Table 2. The corresponding brand name or generic name for each drug is shown in parentheses in the first column. Lower numbers indicate higher degrees of similarity. KL divergence is not symmetric, so we compare drugs D_1 and D_2 by taking

	Allergies	Breast cancer	Flu	Stroke
Symptoms	drowsy (0.24)	tumor (0.09)	cough (0.05)	confused (0.06)
	migraine (0.04)	nasal polyps (0.03)	fever (0.05)	sleepy (0.05)
	headache (0.04)	polyps (0.03)	confused (0.05)	shoulder pain (0.05)
	confused (0.04)	confused (0.03)	throw up (0.04)	headache (0.05)
	sleepy (0.03)	swelling (0.03)	headache (0.04)	throw up (0.04)
	sneeze (0.03)	headache (0.02)	sleepy (0.03)	pass out (0.02)
	swollen (0.02)	lose weight (0.02)	asthma (0.03)	feel sick (0.02)
	asthma (0.02)	fever (0.02)	feel sick (0.02)	cough (0.02)
	cough (0.02)	sleepy (0.02)	pass out (0.02)	shaking (0.02)
	throw up (0.02)	swollen (0.02)	migraine (0.02)	hemorrhage (0.02)
Drugs	zyrtec (0.16)	alcohol (0.14)	alcohol (0.14)	alcohol (0.13)
	alcohol (0.14)	cocaine (0.05)	cocaine (0.06)	aota (0.10)
	adderall (0.03)	tamoxifen (0.03)	prozac (0.03)	cocaine (0.05)
	cocaine (0.03)	aspirin (0.02)	tylenol (0.03)	zofran (0.05)
	prednisone (0.03)	testosterone (0.02)	dopamine (0.02)	testosterone (0.04)
	benadryl (0.02)	histamine (0.02)	adderall (0.02)	actos (0.02)
	morphine (0.02)	amoxicillin (0.02)	prednisone (0.02)	adderall (0.02)
	imitrex (0.02)	metformin (0.01)	viagra (0.02)	risperdal (0.01)
	valium (0.01)	clarithromycin (0.01)	risperdal (0.02)	viagra (0.01)
	histamine (0.01)	vitamin c (0.01)	benadryl (0.02)	codeine (0.01)

Table 1: The symptoms and drugs most strongly associated with four common conditions. Many symptoms (e.g., migraine, headache, sneeze, etc.) and drugs (e.g., zyrtec, prednisone, benadryl) are correctly associated with allergy condition. The other conditions are correctly associated with many symptoms, but not with many drugs.

	Acetaminophen (Tylenol)	Advil (Ibuprofen)	Aleve (Naproxen)	Aspirin (no equivalent)	Ibuprofen (Advil)	Naproxen (Aleve)	Tylenol (Acetaminophen)
Acetamin.	-	3.46	3.49	4.80	3.19	4.68	1.77
Advil	3.46	-	0.66	3.94	0.32	1.11	1.81
Aleve	3.49	0.66	-	4.91	0.54	2.03	2.44
Aspirin	4.80	3.94	4.91	-	4.23	5.24	2.90
Ibuprofen	3.19	0.32	0.54	4.23	-	1.37	1.84
Naproxen	4.68	1.11	2.04	5.24	1.37	-	3.33
Tylenol	1.77	1.81	2.44	2.90	1.84	3.33	-

Table 2: KL divergences between nonsteroidal anti-inflammatory drugs derived from drug-symptom distributions. Lower numbers indicate a higher similarity. Each drug's brand name and generic name was treated as a unique drug for the purpose of evaluating the drug similarities. Our model correctly identifies Acetaminophen and Tylenol and Ibuprofen and Advil as similar drugs, but fails to identify Naproxen and Aleve as being similar. The relative results do not change when drug-condition distributions are instead used to compute the similarity.

the mean of the KL divergences of their drug-symptom distributions. This approach can also be used to measure the similarity between distributions conditioned on two conditions or two symptoms.

Our model correctly indicates that Ibuprofen is the most similar drug to Advil and that Acetaminophen is the most similar drug to Tylenol. It incorrectly indicates that Naproxen and Aleve are more similar to Ibuprofen and Advil than the two drugs are to each other. This error may be caused by a much lower number of tweets mentioning Aleve and Naproxen; these drug terms occur in our corpus approximately 20% and 8% as often as the next most infrequent term (Acetaminophen), respectively.

Condition symptoms vs. drug side effects

We evaluate how well our model can be used to distinguish symptoms caused by a condition from symptoms caused by a drug (i.e., drug side effects) by comparing conditional probabilities as shown in Eq. 4. The drugs and conditions to compare were chosen by selecting the five most frequently occurring drugs with a clearly associated condition; some drugs such as morphine, adderall, aspirin, and benadryl occurred more frequently, but were not strongly associated with any condition (as determined by $\Pr C|D$). Such drugs either belonged to more general classes of drugs that are often used for symptom relief (i.e., NSAIDs and antihistamines) or were drugs that are known to be commonly abused (e.g., morphine, adderall, xanax, etc.). The top five drugs and their associated conditions are: prednisone (allergies), lipitor (diabetes), prozac (depression), zoloft (depression), and paxil (depression).

The top ten symptoms attributed to each drug and condition are shown in Table 3. Symptoms attributed to the drug are shown in the D rows (top half) and symptoms attributed to the condition are shown in the C rows (bottom half). The symptoms associated with depression (D=Prozac,

	D=Prednisone, C=Allergies	D=Lipitor, C=Diabetes	D=Prozac, C=Depression	D=Zoloft, C=Depression	D=Paxil, C=Depression
D	migraine (.15)	headache (.05)	caries (.06)	pulmonary emb. (.03)	high BP (.08)
	swollen (.05)	stomach ache (.02)	high BP (.05)	embolism (.03)	suicidal th. (.06)
	feel sick (.05)	migraine (.02)	suicidal th. (.04)	gynecomastia (.03)	clubfoot (.06)
	mood swings (.02)	liver damage (.02)	tooth decay (.04)	deep vein thromb. (.03)	gain weight (.06)
	shaking (.02)	deep vein thromb. (.02)	gain weight (.03)	sneeze (.03)	tooth decay (.04)
	wheezing (.02)	pulmonary emb. (.02)	clubfoot (.02)	tumor (.02)	caries (.03)
	swelling (.02)	blood clot (.02)	irritability (.02)	suicidal th. (.01)	urinary incont. (.03)
	exhausted (.01)	gain weight (.01)	low testosterone (.02)	high BP (.01)	incont. (.03)
	asthma attack (.01)	muscle soreness (.01)	irritable bowel (.02)	blood clot (.01)	diarrhea (.02)
	throw up (.01)	suicidal th. (.01)	inflammation (.02)	vein thromb. (.01)	irritability (.02)
C	drowsy (.24)	high BP (.02)	confused (.06)	sleepy (.03)	confused (.06)
	sneeze (.02)	confused (.01)	sleepy (.04)	confused (.03)	sleepy (.05)
	headache (.02)	lose weight (.01)	throw up (.03)	lose weight (.02)	throw up (.03)
	cough (.02)	cough (.01)	cough (.03)	pass out (.02)	cough (.03)
	polyps (.02)	asthma (.01)	pass out (.02)	throw up (.02)	pass out (.03)
	hangover (.01)	sleepy (.01)	headache (.02)	exhausted (.01)	headache (.02)
	tumor (.01)	pass out (.01)	exhausted (.02)	cough (.01)	exhausted (.02)
	sleepy (.01)	exhausted (.00)	shaking (.02)	hangover (.01)	shaking (.02)
	fever (.01)	throw up (.00)	lose weight (.02)	cramps (.01)	feel sick (.02)
	runny nose (.01)	feel sick (.00)	cramps (.01)	inflammation (.01)	lose weight (.01)

Table 3: Symptoms most strongly associated with drugs (D row) and conditions (C row) for each drug and condition pair. Many drug side effects are correctly identified (e.g., nausea with Prednisone, weight gain with Prozac and Paxil, etc.). Similarly, there is high agreement among the conditions associated with depression with no more than two entries differing between any pair. Note the following terms were abbreviated: *embolism*, *high blood pressure*, *incontinence*, *suicidal thoughts*, and *thrombosis*.

D=Zoloft, and D=Paxil) are strikingly similar, with only one entry differing between the Prozac and Paxil columns (i.e., cramps vs. feel sick) and two entries unique to the Zoloft column (i.e., hangover and inflammation). The symptoms associated with the drugs that treat depression are less accurate, with several terms that do not appear to be related at all (e.g., clubfoot, sneeze, tumor, irritable bowel, etc.). The drug symptoms caries, tooth decay, weight gain, incontinence, and diarrhea are known side effects.

Similarly, many of the symptoms associated with Prednisone (i.e., swelling, mood changes, nausea, and exhaustion) and Lipitor (i.e., headache, migraine, weight gain, muscle soreness, and stomach pain) are known side effects of those drugs. Many of the symptoms associated with allergies are allergy symptoms, such as sneeze, headache, cough, and runny nose, whereas fewer symptoms appear to be correctly associated with diabetes (i.e., exhaustion, weight loss, and nausea). These results illustrate that while we differentiate between symptoms caused by conditions and symptoms caused by drugs (i.e., drug side effects), identifying causal relationships is difficult and should be handled with care.

Conclusion

We described a model for learning associations between mentions of drugs, symptoms, and medical conditions in Twitter, and investigated its ability to *(1)* learn associations between drugs, symptoms, and conditions, *(2)* to identify conditions or drugs that are similar based on their associated symptoms, and *(3)* to differentiate between symptoms caused by drugs (i.e., drug side effects) and symptoms caused by a condition that a drug is being taken to treat.

We find that our approach is often able to correctly identify equivalent drugs as similar and to correctly separate a condition's symptoms from drug side effects. We envision incorporating our approach with health-related text mining systems to improve their accuracy. Systems for discovering expected and unexpected drug side effects could benefit from our method for differentiating between conditions' symptoms and drug side effects, for example, and our drug similarity and condition similarity measures could be used to help identify drug and condition synonyms.

References

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

Chen, L.; Hossain, K. S. M. T.; Butler, P.; Ramakrishnan, N.; and Prakash, B. A. 2014. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, ICDM '14.

Kuhn, M.; Letunic, I.; Jensen, L. J.; and Bork, P. 2016. The sider database of drugs and side effects. *Nucleic Acids Res* 44(Database issue):D1075–D1079.

Paul, M., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *International AAAI Conference on Web and Social Media*.

Yates, A., and Goharian, N. 2013. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. In *Proceedings of the 35th European conference on Advances in Information Retrieval*.

Yates, A.; Goharian, N.; and Frieder, O. 2015. Extracting Adverse Drug Reactions from Social Media. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'15)*.