Detecting Privacy-Sensitive Events in Medical Text

Prateek Jindal
Yahoo! Inc.
2021 S. First St
Suite 110
Champaign, IL, USA
pjindal@yahoo-inc.com

Carl A. Gunter
Department of Computer Science, UIUC
201 N. Goodwin Ave
Urbana, IL, USA
cgunter@illinois.edu

Dan Roth
Department of Computer Science, UIUC
201 N. Goodwin Ave
Urbana, IL, USA
danr@illinois.edu

ABSTRACT

In this paper, we present a novel semi-supervised technique for finding privacy-sensitive events in clinical text. Unlike traditional semi-supervised methods, we do not require large amounts of unannotated data. Instead, our approach relies on information contained in the hierarchical structure of a large medical encyclopedia.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Concept Learning*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*; J.3 [**Life and Medical Sciences**]: Medical Information Systems; K.5.2 [**Legal Aspects of Computing**]: Government Issues—*Regulation*

General Terms

Algorithms, Experimentation, Performance, Security

Keywords

Natural Language Processing, Health Informatics, Computer Security, Active Learning, Semi-Supervision, Set Expansion, Concept Identification, Domain Knowledge, SNOMED CT

1. INTRODUCTION

There is a growing interest in ways to exploit the increased use of Electronic Health Records (EHRs) to improve healthcare and reduce its costs. One key goal is to enable records of patients to be shared quickly to assure that the right information is in the right place when it is needed (such as a list of allergies and medications for a patient who has arrived in an emergency room) and that redundancy is reduced (so that tests run at one institution do not need to be re-run at another). This sharing raises privacy concerns, even when it can be taken as given that a patient wants to allow his records to be shared to aid his own care. One area of concern is the risk of sharing what some patients and regulators view as extra-sensitive information when this information is not needed for a specific instance. For instance, a patient with a twisted ankle would not like her records of childhood abuse to be shared automatically for this type of treatment.

There are several types of sensitive data that are found in the clinical narratives. We categorize the sensitive data into 5 major types below:

1. Mental health and abuse in the family
2. Substance Abuse
3. HIV Data
4. Presence of genetic information in EHRs
5. Sexually transmitted infections and reproductive health

In some hospital systems, structured information in the form of ICD9 codes is used to restrict the sharing of patients' clinical history. For instance, if a record contains ICD9 code 042 (HIV) and a patient has indicated that his HIV status should not be shared, then this can be done by preventing all or some of his record (including the HIV diagnosis) from being shared. This strategy has a number of limitations, one of the most serious of which is the fact that it does not work for *unstructured* data like clinical narratives.

In this paper, we present a semi-supervised technique for finding phrases in clinical narratives which express a sensitive condition about the patient. The advantage of our technique is that we don't require any annotated data for training etc. Moreover, we don't even require any unannotated data. This is in contrast to traditional semi-supervised approaches [9, 6] which rely on large amounts of unannotated data for building distributional contexts. Instead of relying on distributional contexts of the relevant concepts, we use the tree positions of the concepts in a medical encyclopedia named SNOMED CT[1] to find the similarity between them.

Problem of finding sensitive information in clinical notes is complicated by the fact that the concepts may be negated in the text or the concepts may not apply to the patient at all. We briefly discuss these issues as well.

2. DESCRIPTION OF OUR METHOD

The input to our algorithm consists of a few examples (seeds) of the concepts which the user is interested to find. For example, if one wants to find the instances of *substance abuse*, one may give the following concepts in the input: *"cocaine, barbiturates and alcohol"*. Let us represent this

[1]http://www.ihtsdo.org/snomed-ct/

BCB'14, September 20–23, 2014, Newport Beach, CA, USA.
ACM 978-1-4503-2894-4/14/09.
http://dx.doi.org/10.1145/2649387.2662451.

Level	Concepts	
Level 0	Cocaine,	Cocaine measurement
Level 1	Drug measurement, **Psychostimulant,**	Tropane alkaloid, Ester type local anesthetic
Level 2	Azabicyclo compound, Measurement of substance, Psychotherapeutic agent, Heterocyclic compound,	Local anesthetic, Ester, Stimulant, Tropane alkaloid
Level 3	CNS drug, Aza compound, Drug pseudoallergen by function, Measurement,	**Psychoactive substance,** Anesthetic, Alkaloid, Organic compound
Level 4	General drug type, Drug pseudoallergen, Techniques,	Substance categorized functionally, Evaluation procedure, Chemical categorized structurally

Table 1: This table shows the descriptor for concept "cocaine".

seed set by $\mathcal{S}$. Let s_i denote the i^{th} element of this seed set. Now, the various steps of our algorithm are described in the following subsections.

2.1 Step 1: Building Individual Descriptors

Using SNOMED CT, we first build a detailed descriptor of every concept specified in the seed set $\mathcal{S}$. A concept can appear at multiple places in SNOMED CT. We define the descriptor of a concept to be simply the parents of the concept upto 5 higher levels. We explain it below with the help of an example. Let us consider the concept "cocaine". The descriptor of this concept is shown in Table 1. At level 0, two SNOMED CT concepts corresponding to "cocaine" are shown. Concepts at any level $i+1$ are basically the parents of concepts at level i. It is normal for some of the concepts to repeat at later levels. These descriptors were made by a simple breadth-first search on the SNOMED CT graph starting from the concept under consideration.

2.2 Step 2: Obtaining Model for Seed-Set Using Active Learning

After computing the individual descriptors of all the concepts in the seed set $\mathcal{S}$, we merge those descriptors into a single descriptor. Let us assume that for concept x, parents at level i are denoted by the set $\mathcal{L}_i(x)$. Then the levels of the overall descriptor are defined by the following equation:

$$\mathcal{L}_i(\mathcal{S}) = \bigcup_{j=1}^{|\mathcal{S}|} \mathcal{L}_i(s_j) \qquad \forall i \tag{1}$$

After some preprocessing (removal of overly general concepts), the descriptor is shown to the user. Then the user is supposed to identify one or more most appropriate SNOMED CT concepts from this overall descriptor. User response is recorded into a list. This list constitutes the learned model for the concept identification. Let us call this list as $MedRep(\mathcal{S})$. No further input from user is now required. To give an example, in one of our sample runs, $MedRep(\mathcal{S})$ contained the following SNOMED CT concepts for the seed set described at the beginning of this section: *Psychoactive substance, Anxiolytic, sedative AND/OR hypnotic, Sympathomimetic agent, Centrally acting hypotensive agent, Centrally acting muscle relaxant, Morphine Derivative, Opiate, Hallucinogen, Alcoholic Beverage, Central Depressant.*

2.3 Step 3: Computing the Similarity of any Concept to Seed-Set

In this subsection, we describe how to compute the similarity of any given SNOMED CT concept to the seed set, $\mathcal{S}$, provided by the user. Let us denote the given SNOMED CT concept by the variable x. Then the similarity, $sim(x, \mathcal{S})$, of the concept x to the seed set $\mathcal{S}$ is defined by the following equation:

$$sim(x, \mathcal{S}) = \left| \left(\bigcup_{i=1}^{4} \mathcal{L}_i(x) \right) \bigcap MedRep(\mathcal{S}) \right| \tag{2}$$

In other words, similarity of a concept to the seed set is the number of *unique* SNOMED CT concepts in the descriptor of the concept that also appear in the seed-set model learned in step 2 above.

2.4 Step 4: Concept Identification

First of all, we map the noun phrases occurring in the clinical text to corresponding SNOMED CT concepts using MetaMap software[2]. Then, by using the procedure described in Step 3 above, we find the similarities of all the concepts in any given document to the given seed set. Finally, we rank the concepts in decreasing order of similarity and then apply a threshold (which was computed empirically). This gives us the desired list of most relevant concepts in the given document which belong to the same category as the seeds input by the user to begin with.

3. ASSERTION STATUS

In the above section, we described a technique for finding sensitive events in the clinical text. However, it should be appreciated that the mere presence of a sensitive concept in clinical text does not necessarily raise privacy concerns about patient's medical information.

We give two examples to illustrate this fact. For the first example, consider a clinical narrative that says, *"The patient*

[2]http://metamap.nlm.nih.gov/

does not inject heroin". Even though the drug *heroin* is mentioned in this sentence, there is no sensitive information because of the negation. For the second example, consider the sentence, *"In patient's workplace, drug abuse and smoking is very common"*. In this sentence also, there is no sensitive information because the concepts *drug abuse* and *smoking* are not telling anything about the patient himself.

From the above discussion, we can see that it is important to find out the following information about the identified concepts:

1. *Negation Detection*: Whether the concept has been negated or not
2. *Experiencer Detection*: Whether the concept belongs to the patient or not

We implemented a system to find the above 2 attributes of the concepts. Our approach is based on the rule-based method for assertion classifier developed by Xu et al [10]. In our system, we used the lists provided by Xu et al and also the lists contained in an implementation[3] of ConText algorithm [2, 1].

First of all, we identify the trigger words for negation and experiencer detection. Consider the following sentence as an example:

> The patient denies any IV drug use but did describe cocaine use for last 2 months.

In this sentence *'denies'* is the trigger word for negation. It is important to note that our algorithm differentiates between pseudo-triggers (like 'no increase', 'not cause' etc.) and the actual trigger words.

After determining the triggers, we determine the scope of the trigger words. The scope of a trigger word generally starts from the word to the right of the trigger and extends till the end of the sentence. But certain words (like 'but' in the above example) can cause the scope of a trigger to end early. Also, for certain triggers, the scope lies to the left of the trigger instead of the right. For example, consider the following sentence:

> Lung injury was ruled out by the MRI exam.

The scope of *'was ruled out'* is *'Lung injury'*. If a concept falls within the scope of some trigger word, we associate the appropriate attribute to that concept.

The confidence of prediction depends on the distance between the trigger word and the concept. When a concept is close to the trigger word, then it is more likely that the trigger word is associated with the concept.

4. RESULTS

4.1 Concept Identification

We tested the effectiveness of our approach for finding the instances of substance-abuse events in the corpus that we had. Our corpus consists of 2 parts: (a) *i2b2 data*: 50 documents from *Partners HealthCare Institute* which were made available in *2010 i2b2 shared task* and (b) *Web data*: Discussions among individuals collected from online medical forums. We used the seed set described in §2. Using this simple seed set, we were able to identify in our corpus several other products which are often used for substance-abuse. For example, our algorithm revealed the following additional concepts: *amphetamines, heroin, benzodiazepines, marijuana, buprenorphine, butane, hallucinogens, narcotics, wine* etc. Overall, we achieved precision, recall and F1 values of 79.2, 49.1 and 60.6 respectively.

[3]https://code.google.com/p/negex/

4.2 Negation and Experiencer Detection

We tested our approach on the i2b2 portion of our dataset. We achieved the F1 scores of 78.2 and 79.6 respectively for the tasks of negation and experiencer detection.

5. ERROR ANALYSIS

5.1 Concept Identification

We noted above that our system has a recall of 49.1 for the task of concept identification. The relatively lower recall value indicates that our system still misses many drugs that are used for abuse. Table 2 shows some of the drugs that were missed by our system. There were three main reasons for missing drugs: *trademark names, street names and abbreviations*. We briefly describe these three factors below.

SNOMED CT does not always have the trademark names for the drugs. For example, *Lorazepam* is a drug that can potentially be abused. Its tradename is *Ativan*. Although, SNOMED CT has an entry for *Lorazepam*, it does not have an entry for *Ativan*. Similar thing happened with the concepts *Percocet*, *Vicodin*, *Darvocet*, *Ritalin* and *Lorcet* which were tradenames for *oxycodone*, *hydrocodone*, *propoxyphene*, *methylphenidate* and *hydrocodone bitartrate* respectively.

Sometimes the drugs are mentioned by their street names which are not present in SNOMED CT. For example, street names for the drug *marijuana* are *ganja*, *grass*, *green*, *Mary Jane* etc. Similarly, street names for the drug *cocaine* are *candy*, *Charlie*, *toot*, *crack* etc.

SNOMED CT sometimes does not have the abbreviations for the drug names. For example, it does not have the abbreviations LAAM (levacetylmethadol), PCP (phencyclidine), and many others.

5.2 Negation and Experiencer Detection

For negation and experiencer detection, we made mistakes on cases which are particularly difficult. For example, consider the following sentence:

> The patient's primary care provider was called to discuss outpatient plans to help the patient stop smoking.

The phrase '*patient stop smoking*' can mislead the system to predict a negated event. However, when we see the overall context, we can see that the patient is still continuing with his/her smoking habit.

Next, consider the following sentence:

> He works as a counselor at an alcohol and drug treatment facility for teenagers.

The word '*alcohol*' can mislead the system to predict a positive drug-abuse event. However, there is no drug-abuse (either positive or negative) being reported here at all.

Missed Concepts							
actiq	adderall	ambien	amytal	zydone	sublimaze	revia	orlaam
anexsia	antabuse	ativan	avinza	xodol	stilnox	ritalin	oramorph
biocodone	campral	concerta	damason-P	xanax	steroids	symtan	PCP
darvocet	darvon	demerol	depade	vivitrol	speed	quaalude	roxanol
desoxyn	dexedrine	dextrostat	di-gesic	vicoprofen	suboxone	percodan	tramal
dicodid	dilaudid	duodin	duragesic	vicodin	soma	percocet	norco
duramorph	fioricet	fiorinal	halcion	valium	seconal	panacet	nembutal
hycodan	hydrococet	kadian	kapanol	ultram	ryzolt	rohypnol	mushrooms
klonopin	LAAM	librium	lorcet	tylox	roxicodone	temesta	methadrine
lortab	luminal	ms contin	msir	tussionex	subutex	palladone	

Table 2: This table shows some of the concepts that our algorithm failed to detect.

6. FUTURE WORK

Following are some of the directions for future work:

1. Wikipedia can tell the tradenames, street names and common abbreviations for a lot of drugs. So, a good direction for future work is to extract the medical knowledge in Wikipedia and put it in a structured database.
2. Like Wikipedia, there are several other sources of medical information on the web such as MedlinePlus (online information service produced by the United States National Library of Medicine) and MediLexicon (which gives a lot of useful medical abbreviations). Such sources can be integrated into our system.
3. Another good way to get useful medical knowledge is to send automated queries to web search engines. The top pages from the search results can then be used to glean useful medical information.

7. RELATED WORK

Uzuner et al. [8] present i2b2 shared task on extracting concepts from clinical notes. There are two fundamental differences between i2b2 shared task and our approach. Firstly, i2b2 shared task focused on concepts like tests, treatments and problems and did not really discuss the sensitive information contained in clinical notes. Secondly, unlike our approach, i2b2 shared task primarily focused on supervised approaches. The problem of semi-supervised concept identification is very similar to the problem of set expansion. Previously, we showed the use of negative examples for the set-expansion problem [3]. Usage of negative examples can be quite helpful for the task presented in this paper as well. Finally, we would like to point out some of the works where Wikipedia has been used as a source of medical knowledge: Concept Identification [7, 5], Coreference Resolution [4] and Relation Identification [10].

8. CONCLUSIONS

In this paper, we presented a semi-supervised technique for finding privacy-sensitive information in clinical notes. We showed that even if we start with a small seed-set, our approach is very effective to find out the concepts of interest. We also discussed the issues related to negation and experiencer detection.

9. ACKNOWLEDGMENTS

We thankfully acknowledge the contribution of Tony Michalos in helping us to prepare the datasets. This research was supported by Grant HHS 90TR0003/01. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the HHS or the US government.

10. REFERENCES

[1] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *JBI*, 34(5), 2001.

[2] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *JBI*, 42(5), 2009.

[3] P. Jindal and D. Roth. Learning from negative examples in set-expansion. In *Proceedings of the 2011 Conference on International Conference on Data Mining (ICDM)*, pages 1110–1115. IEEE Computer Society, 2011.

[4] P. Jindal and D. Roth. End-to-end coreference resolution for clinical narratives. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI)*, pages 2106–2112. AAAI Press, 2013.

[5] P. Jindal and D. Roth. Using soft constraints in joint inference for clinical concept recognition. In *EMNLP*, pages 1808–1814, 2013.

[6] P. Pantel, E. Crestan, A. Borkovsky, A. Popescu, and V. Vyas. Web-scale distributional similarity and entity set expansion. In *EMNLP*, 2009.

[7] K. Roberts and S. Harabagiu. A flexible framework for deriving assertions from electronic medical records. *JAMIA*, 18(5), 2011.

[8] O. Uzuner, B. South, S. Shen, and S. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *JAMIA*, 2011.

[9] R. Wang and W. Cohen. Language-independent set expansion of named entities using the web. In *ICDM*, pages 342–350. IEEE Computer Society, 2007.

[10] Y. Xu, K. Hong, J. Tsujii, I. Eric, and C. Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *JAMIA*, 19(5):824–832, 2012.