# Building a Patient-based Ontology for User-written Web Messages

**Marina Sokolova**
Faculty of Medicine,
University of Ottawa
and
Electronic Health Information Lab,
CHEO Research Institute
sokolova@uottawa.ca

**David Schramm**
Faculty of Medicine,
University of Ottawa
and
Children's Hospital of Eastern Ontario
The Ottawa Hospital
dschramm@toh.on.ca

## Abstract

We introduce an ontology that is representative of health discussions and vocabulary used by the general public. The ontology structure is built upon general categories of information that patients use when describing their health in clinical encounters. The pilot study shows that the general structure makes the ontology useful in text mining of social networking web sites.

## 1 Introduction

Recent studies have shown that public health surveillance benefits from information posted by users on the Web (Carneiro and Mylonakis, 2009; Ginsberg et al, 2008). Health-related messages can be found on Web forums hosting social networks (e.g., www.PatientsLikeMe.com) or individual blogs (e.g., http://www.jackslemonade.com).

For medical professionals, the user-written health information assists in prediction of public attitude towards health policies. In user messages, patient-based information prevails over biomedical information. Patient-based information is brought forth when a user views himself as a potential or real patient of a health care provider. This information reveals details of one's health that are usually discussed during visits to a health care provider. Patient-based information is often identified as evidence-based, whereas the biomedical information is viewed as knowledge-based (Hersh, 2009).

Development of social media has prompted refocusing of text analysis from biomedical to patient-based health information mining. Several academic groups actively work on health information studies (Angelova, 2010; Chapman, 2010; Chanlekha and Collier, 2010). These groups work on methods for the analysis of academic and professional articles in medical journals and traditional news media, as well as hospital documentation.

At present, user-written health information is the subject of studies by data mining where the analysis primarily relies on statistical methods (Lampos and Christianini, 2010) and public health informatics which usually addresses specific questions, e.g., injury discussions by military servicemen (Konovalov et al, 2010).

A prevalent trend in health-related text analysis is to solve a particular task which is closely associated with a particular data source, e.g., identifying involuntary childlessness terminology on a dedicated web site (Himmel et al., 2009) or finding new terms used on a patient social networking site (Doing-Harris and Zeng-Treiler, 2011). The specific focus makes the accumulated knowledge inherently individualized towards the task and the data. It permits high accuracy on the original data, whereas shifting to other data sets is likely to experience performance set-back.

Our goal is to build a patient-based resource organized as an ontology, a repository of health-related terms assigned into a hierarchical structure of semantic categories. The general categories are durable and able to withstand the rapidly evolving environment of the Web. In an empirical setting, we show that the ontology content is representative of health-related topics and vocabulary used by the general public on the Web.

## 2 Motivation

Major health concerns, related events and issues, and behavioural trends can be identified from what people post on social networks. The importance of this analysis became more pronounced during the H1N1 pandemic as recent research demonstrates (Lampos and Christianini, 2010).

User-written health information extraction can be challenging in a two-fold way:



*Proceedings of Recent Advances in Natural Language Processing*, pages 758–763,
Hissar, Bulgaria, 12-14 September 2011.

| Twitter |
|---|
| 11: i can't cos i haven't slept yet and it's 9:43am. i'm having some serious insomnia. i'm trying to sleep but i keep checking mail.<br>12: the doctor came, examined me and told me i had early tonsilitis. will look it up on the net. i'm in my mom's room while my room aerates. |
| **20 News groups** |
| I sometimes see OTC preparations for muscle aches/back aches that combine aspirin with a diuretic. The idea seems to be to reduce inflammation by getting rid of fluid. Does this actually work? |
| **MySpace** |
| i thouroughly understand ur point though, my grandmother has lung cancer so i cant stand smoking, its all a personal choice; you cant change someones mind if they choose not to listen. . . . |
| **Amazon.com** |
| Just purchased this blender & am returning it immediately. It has a number of terrible features: it's very difficult to remove the cover if you have carpal tunnel, arthritis, or weak hands. |

Figure 1: Examples of user messages.

i various web sites host texts written in different styles (Figure 1 lists samples from four web sites); thus, a site-specific method has an application range limited to the site;

ii existing text mining tools focus on biomedical and professional terminology that may be absent in social media (Casoto et al, 2010); as a result, these tools need a considerable readjustment before application to user-written text.

Standardized classification of diseases and other health-related problems is critical for epidemiologic and health management purposes. At the same time, there are few publications dedicated to user-written health information. In one study (Doing-Harris and Zeng-Treiler, 2011), the authors looked for new health-related terms in messages posted on `PatientsLikeMe.com`. User requests posted on an involuntary childlessness message board were studied in (Himmel et al., 2009). Blogs written by military servicemen were studied by (Konovalov et al, 2010). The researchers sought terms that described clinically relevant combat exposure. All the three listed studies have a restricted appeal: each was carried out on one data set only and was not applied or reproduced on other data sets.

Biomedical information extraction and text classification have a successful history of method and tool development, including deployed information retrieval systems (Hersh, 2009), knowledge resources and ontologies (Cohen et al, 2010; Yu, 2006). Exponential increase in bio-, bioinformatics and medical publications has caused a rapid development of ontologies that help to recognize and categorize research and professional vocabulary (Yu, 2006). We discuss here a few examples.

*GENIA*[1] is built for the microbiology domain. Categories include DNA-metabolism, Protein-metabolism, and Cellular process. *Medical Subjects Heading (MeSH)* is a controlled vocabulary thesaurus, produced by the National Library of Medicine[2]. Its terms are informative to experts but might not be in use by the general public (e.g., Work Schedule Tolerance at the top level and Motor Cortex, Trypanosoma cruzi at the bottom level). The *Medical Entities Dictionary (MED)*[3] is an ontology containing approximately $60,000$ concepts, $208,000$ synonyms, and $84,000$ hierarchies. This powerful lexical and knowledge resource is designed with medical research vocabulary in mind. *Unified Medical Language System (UMLS)* has 135 semantic types and 54 relations that include organisms, anatomical structures, biological functions, chemicals, etc.

Another internationally recognized classification scheme is the Systematized Nomenclature of Medicine Clinical Terms (*SNOMED CT*) maintained by the International Health Terminology Standards Development Organization.[4] Although *SNOWMED CT* is considered to be the most comprehensive clinical health care terminology classification system, it is primarily used to permit standardization of electronic medical records rather than to mine user-written health-related content. A public health ontology *BioCaster*[5] is built for surveillance of traditional media. It helps to find disease outbreaks and predict possible epidemic threats.

[1] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html`

[2] `http://www.nlm.nih.gov/mesh/`

[3] `http://med.dmi.columbia.edu/`

[4] `http://www.nlm.nih.gov/research/umls/Snomed/snomed$\_$faq.html` Accessed 18/07/2011

[5] `http://born.nii.ac.jp/?page=ontology`

All these sources would require considerable modification before they could be used for analysis of messages posted on public Web forums.

## 3 Methodology

Adequate patient treatment depends on a correct understanding of what people say about their health and cross-referencing of the terms they use (Aspden et al., 2003). We began by building a set of semantic categories that a patient would use when discussing personal health in a clinical setting.

There are several internationally accepted interrelated disease and health-related problems classification schemes:

- The International Statistical Classification of Diseases and Related Health Problems (ICD-10) developed by The World Health Organization is the internationally recognized standard diagnostic classification system (ICD–10, 2004).

- The International Classification of Procedures in Medicine (ICPM) categorizes medical and surgical procedures (ICPM, 1978).

- The International Classification of Functioning, Disability and Health (ICF) categorizes and qualifies disability, physiological functioning of body systems and their impairment, anatomical parts of the body and their impairment, activities of an individual and their limitations, participation in life situations and their restrictions, and health-related environmental factors (ICF, 2001).

We amalgamated and streamlined these international health related classification scheme taxonomies to facilitate the classification of user-written health-related content on the web. Extensive clinical experience of one of the authors was applied to empirically adapt the classification scheme to users' description of their health on various social networking web sites. Figure 2 shows the ontology structure.

- Person
  - Anatomical parts of the body
  - Physiological functioning of body
- Diseases and Health-Related Problems
  - Diseases
  - Symptoms
- Health Care System
  - Health Care Providers
    - Physician
    - Nurse
    - Physiotherapist
    - Psychologist
    - Other Health Care Provider
  - Health Care Setting
    - Hospital
    - Ambulatory surgery center
    - Physician Office
    - Community Health Care Clinic
    - Other Health Care Setting
  - Health Care Procedures
    - Diagnostic
    - Therapeutic
- Health-Related Environmental Factors
  - Physical
  - Social
  - Attitudinal

Figure 2: The structure of the ontology.

We populate the categories with terms found in sources that provide patient-friendly terminology.[6] Many of the terms utilized in the International Classification of Functioning, Disability and Health (ICF) and International Statistical Classification of Diseases and Related Health Problems (ICD-10) nomenclature are typical of the vocabulary that individuals may use to describe their health related states (adapted for Diseases and Symptoms subcategories).

We used Merriam-Webster Visual Dictionary to add the Person terms and Webster's New World Medical Dictionary to the Procedures subcategory. Provider and Setting terms were adapted from lists of certified medical doctor boards [7] and associations of other health occupations[8].

[6]The ontology is posted on http://www.ehealthinformation.ca/ap0/opendata.asp.

[7]http://www.certificationmatters.org/about-board-certified-doctors/

[8]http://www.ama-assn.org/ama/pub/

# 4 Data

To assess the ontology usefulness, we used publicly available data sets *20 News Groups*[9], *Twitter* and *MySpace* [10], and *Amazon.com*[11].

**20 News groups** has $20,000$ texts divided into 20 groups, including a group of medical texts. The medical group consists of 990 messages gathered from Web chat boards. In these messages, users discuss their health problems, ask questions pertaining to health, give advice and share relevant experience. The set has $239,120$ words, an average length of a message is $242$ words, including partial citations of previous messages when applicable. Full grammatical constructs and a rich lexicon make the messages reminiscent of a more traditional, pre-Internet writing.

**Twitter** is a micro-blogging service, with instant message postings. It is organized as a social network of Twitter users. A user can post short messages, no longer than 140 characters, that are publicly visible by default. Other users can subscribe to these tweets (i.e., become followers) and respond with their messages. A user can group his messages by topic or types and make them accessible only to followers. URL shortening is common, e.g., *goo.gl* for *www.google.com*. Other condensing happens through shorthand (e.g., "LOL" (laugh out loud),"DWT" (driving while texting), "4gt"(forgot)) and emoticons (e.g., `;-)` ).

We worked with $30,164$ threads of consequent tweets. The treads are split into two subsets, $3,754,668$ and $15,199,470$ words. An average length of a thread is 560 words, albeit some words can be very short (e.g., "u","4").

**MySpace** (aka My___, Myspace) has been a leading social network in 2006 – 2008, when 95 million unique users visited the web site in a year. Friends can leave their comments in the user's "Friends Space"; it is left to the user's discretion to keep or delete those comments or mandate to approve them before posting. Users can assign emoticons to posts (e.g., `:-0`, `:-(`). Ability to reach all friends simultaneously is given through bulletins, messages posted on the bulletin board and remaining there for 10 days. Profiles have enhanced blogging that promotes longer posts. However, a typical post may exhibit the Internetspeak features, such as the the shorthand and simplified grammar (e.g., "l8r" (later), "c u" (see you) ).

We analyzed $18,178$ posts split into four subsets of $218,628$, $1,219,730$, $1,987,495$ and $9,403,345$ words respectively. An average length of a post available to us is $167$ words.

**Amazon.com** posts user reviews of consumer products. In the reviews, users share their experience and opinion about the products. Those comments are often accompanied or illustrated by a narrative of real life events included health-related problems. Messages are organized according to the types of the assessed goods.

We worked with $8,000$ reviews, evenly split along four topics: books ($349,530$ words) , DVD ($337,473$ words), electronics ($222,862$ words), and kitchen&houseware ($188,137$ words). An average length of reviews is counted in words: books – $175$, DVD – $169$, Electronics – $111$, Kitchen – $99$. The grammatical structure and vocabulary are rich enough to provide meaningful communication and lexical information.

# 5 Empirical Results

We built $N$-gram word models ($N = 1,2,3,4$). The $N$-gram models estimate the probability of a word sequence $w_1 \ldots w_n$ appearing in the data. The estimate is computed as a conditional probability of the word $w_n$ appearing after the sequence of words $w_1 \ldots w_{n-1}$:

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-N+1}^{n-1})s \qquad (1)$$

We searched all four data sets for the presence of the ontology terms. In each data set, we concentrated on terms with occurrence $\geq 10$. These words are more likely to be representative across many users, but not indicative of individual preferences. Representativeness of the ontology categories varied in coverage and support. Within the data sets, *Body* and *Symptoms* were represented by 80% – 90% of their terms, a larger proportion than other categories. Although only $30\%$ – $50\%$ of *Doctor* terms were extracted from every data set, the found terms were among the most frequent in every corpora (e.g., *doctor*, *physician*).

The term disambiguation was especially important for non-professional terms which could have

`education-careers`

[9] `http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html`

[10] `http://caw2.barcelonamedia.org/node/7`

[11] `http://www.cs.jhu.edu/~mdredze/datasets/sentiment/`

| Relevance | Data | Post |
|---|---|---|
| Relevant | Twitter | thinkin there's a doctor's appointment in my future. tired of being sick. need to get back to taking care of my family before christmas. |
| | MySpace | my best friend stephanie's brother mike's best friend paolo was just diagnosed with a.l.l. leukemia. |
| Irrelevant | MySpace | to ensure the protection of military and civilian personnel in the department of defense from an influenza pandemic, including an avian influenza pandemic. |
| | Amazon | With one hand, pull the superoposterior part of the pinna in a superoposterior direction while inserting the earphone with the other. This straightens the ear canal and makes it easier to insert the earphone. (Your doctor uses the same maneuver when he/she examines you with an otoscope.). |

Table 1: Examples of posts extracted with the health ontology.

| Category | 20 New groups | Twitter | MySpace | Amazon.com |
|---|---|---|---|---|
| Doctors | doctor, physician, radiologist | cardiologist, dermatologist, doctor, gynecologist, pediatrician | cardiologist, doctor, gynecologist, neurologist, pathologist | cardiologist, gynecologist, pediatrician, physician |
| Procedures | diet, circumcision, needles, ultrasound | diet, ecg, homeopathy, massage, pacemaker | abort, colonoscopy, ct, diet | diet, massage, pacemaker, scan, shots |

Table 2: The least ambiguous ontology categories and examples of their terms.

several non-medical meanings (e.g., *head*, *leg*, *assistant*, *lab*). For terms with multiple meanings, corresponding personal pronouns were strong indicators of a reference to individuals (e.g., *my neck* vs. *attachable neck*, *our doctor* vs. *spin doctor*). Tri- and quadri-grams were useful in finding idiomatic expressions that use ontology terms figuratively (e.g., *technophobes won't have a heart-attack*).

To validate our term choice, we manually examined the use of frequent terms in posts. For each term, we randomly selected 3–6 posts in each data set. We then classified the posts as relevant or irrelevant to person's health information. The examined *20 NewsGroups, Twitter, MySpace* posts were relevant, albeit one was an official document on influenza prevention in military. *Amazon.com* presented an example of a difficult data, where many posts were "false positive", i.e., they used health-related terms in a different context. Table 1 lists the post examples. *Doctors* and *Procedures* terms are the least ambiguous and the most effective in identifying patient-oriented information (Table 2).

## 6 Discussion

We have addressed an important issue of tracking health-related information posted by users on the Web. This information is in demand by health care policy-makers, population and community health organizations and medical practitioners.

Information retrieval/extraction and text mining are popular topics in Health Informatics. The field, however, only recently started to investigate health information in user-written texts. Relationship between self-disclosure and stigmatized health conditions in medical information search have been analyzed (Buchanan et al, 2007). Health information disseminated through medical and military blogs have been studied (Lagu et al, 2008; Konovalov et al, 2010).

Topic classification of user-written health messages has been a focus of research (Frank and Bouckaert, 2006). The study aimed to discriminate between messages with different health topics. Our goal is to extract health-related information from messages. When text data mining systems are deployed to analyze health information they often process institutional documents (Angelova, 2010; Cohen et al, 2010; Chapman, 2010; Ware et al, 2009). We instead work with health-related information.

## 7 Conclusions

Our goal is to assist medical practitioners and researchers in the analysis of Web-based social media. For example, medical professionals may wish to follow the understanding in the general population of a common medical condition such as otitis media and the indications for surgical intervention. We designed a set of semantic categories based on international classification schemes and extensive clinical experience of one of the authors. The categories are representative of notions and concepts that patients invoke in presentation of their health in clinical settings. To find adequate terms, we directly accessed clinical resources used by health care practitioners.

The evidence of ontology usefulness has been obtained from social networking sites. The ontology can be further used for detection of posted confidential health information; aggregation of user health concerns within a certain geographic area; survey of public awareness about particular issues. Additionally, the ontology can be used by tools that analyze health information on electronic media other than the Web (El Emam et al, 2010).

## Acknowledgements

This work is in part funded by a NSERC Discovery grant available to the first author and The Ottawa Hospital Academic Medical Organization – to the second author.

## References

Angelova, G. Use of Domain Knowledge in the Automatic Extraction of Structured Representations from Patient-Related Texts. *Procedings of ICCS 2010*, p.p. 14–27

Aspden, P., J. Corrigan, J. Wolcott, S. Erickson (Eds) *Patient Safety: Achieving a New Standard for Care.* Board on Health Care Services, Institute of Medicine, 2003.

Buchanan, T., A. Joinson, C. Paine, U.-D. Reips. "Looking for medical information on the Internet: self-disclosure, privacy and trust", *He@lth Information on the Internet*, **58**: 8–9, 2007.

Carneiro, H. and E. Mylonakis. "Google trends: a web-based tool for real-time surveillance of disease outbreaks", *Clinical Infectious Diseases*, **49**(10), 1557–1564 , Oxford, 2009.

Casoto, P., A. Dattolo, P. Omero, N. Pudota, C. Tasso. "Accessing, Analysing, and Extracting Information from User Generated Contents", in *Handbook of Research on Web 2.0, 3.0, and X.0*, S. Murugesan (ed.), p.p. 312–328, IGI Global, 2010.

Chanlekha, H. and Collier, N. Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports, *Journal of Biomedical Semantics*, **1**(3), 2010.

Chapman, W. A Hybrid Information Model to Represent Clinical and Linguistic Data Extracted from Clinical Narrative Documents. *Proceeding of American Medical Informatics Association*, 2010.

Cohen, K., C. Roeder, W. Baumgartner Jr., L. Hunter, and K. Verspoor Test suite design for biomedical ontology concept recognition systems. *Language Resources and Evaluation Conference*, pp. 441–446, 2010.

Doing-Harris K, Zeng-Treiler Q. Computer-Assisted Update of a Consumer Health Vocabulary Through Mining of Social Network Data. *Journal of Medical Internet Research*, 2011; 13(2):e37.

El Emam, K., Neri, E., Jonker, E., Sokolova, M., Peyton, L, Neisa, A., and Scassa, T.. "The Inadvertent Disclosure of Personal Health Information through Peer-to-peer File Sharing Programs", JAMIA, **17**: 148–158, 2010.

Frank, E., and Bouckaert, R. Naive bayes for text classication with unbalanced classes. *Proceedings of PKDD 2006*, 503–510, 2006.

Ginsberg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, L. Brilliant. Detecting influenza epidemic using search engine query data. *Nature*, **457** (7232), 1012–1014, 2008.

Hersh W., *Information retrieval: a health and biomedical perspective*, 3rd ed., 2009: Springer.

Himmel W, Reincke U, Michelmann H. Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums. *Journal of Medical Internet Research*, 2009; 11(3):e25.

*ICD-10: International Statistical Classification of Diseases and Related Health Problems: tenth revision*, 2nd ed., World Health Organization, 2004.

*International Classification of Functioning, Disability and Health (ICF)*, World Health Organization, 2001.

*International Classification of Procedures in Medicine (ICPM)*, **1–2**, World Health Organization, 1978.

Konovalov S, M. Scotch, L. Post, C. Brandt. Biomedical Informatics Techniques for Processing and Analyzing Web Blogs of Military Service Members, *Journal of Medical Internet Research*, **12**(4), 2010.

Lagu, T., E. Kaufman, D. Asch, and K. Armstrong. 2008. Content of Weblogs Written by Health Professionals. *Journal of General Internal Medicine*, **23** (10): 1642–1646, 2008.

Lampos, V. and N. Christianini. "Tracking the flu pandemic by monitoring the social web". *2nd Workshop on Cognitive Information Processing*, 2010.

Ware, H., C. Mullett, and V. Jagannathan. Natural Language Processing (NLP) Framework to Assess Clinical Conditions. *JAMIA*,**16**:585–589, 2009.

Yu, A. Methods in biomedical ontology. *Journal of Biomedical Informatics*, **39**, 252–266, 2006.