MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-995

Exploring Relations among Semantic Groups: A Comparison of Concept Co-occurrence in Biomedical Sources

Sasikiran Kandula, Qing Zeng-Treitler

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

It has been observed and reported that the patient's mental model of the medical domain is different from that of a health professional and this difference is one of the primary obstacle in the effective communication of health information to patients. In this study, to better understand these mental models, we explored the relations among different semantic groups of concepts in consumer- and professional-generated health content by analyzing concept co-occurrence information in three biomedical sources. We found significant differences in the prevalence of the semantic groups and the strength of co-occurrences between semantic groups in the three sources. The co-occurrence defined by consumers differs from that defined by professionals. The two professional sources have noticeable differences with each other as well. We believe that addressing these differences can help us generate more informative and consumer-friendly health content as well as develop better consumer health informatics applications.

Keywords:

Mental models, Concept co-occurrence, UMLS semantic groups

Introduction

The past decade has seen an exponential increase in the development of health content and informatics applications for consumers. From web sites to search engines to decision support tools to personal health records, a common theme of the applications is to provide consumers with useful information and help them utilize the information to improve their health outcome. However, many obstacles remain: for example, it is recognized that most health information that is accessible to patients in the US is too hard to understand [1, 2]. While millions of consumer search for health information online, their queries are not always efficient or effective [3-5].

To improve consumer health content and applications, prior research has suggested that we need to understand the gap between lay consumers' and healthcare professionals' mental models [5-8]. A better understanding of the lay consumer mental models could help us anticipate consumers' information needs, organize the content being presented to consumers, explain concepts to consumers in a question-answer system or through the process of text simplification, and to provide decision support at the point of need. A related study we have been involved with is the development of a consumer health vocabulary that suggests consumer-friendly synonyms for difficult medical terms [9, 10].

In this study, we examine the differences between a layperson's and a professional's mental models through the use of co-occurrence information. We believe that the co-occurrence of two terms or concepts in a textual artifact indicates the belief of the creator of the artifact that these are related though the exact relation may not be known and by analyzing co-occurrence information we can understand the mental models of the content creators. Here we use the term *artifact* quite broadly to include a wide range of health content types such as consumer health education articles, medical records or a series of terms queried on a search engine.

In the following sections, we analyze and compare the co-occurrence information in three (two professional-generated and one consumer-generated) biomedical sources – a repository of clinical reports, a collection of biomedical journal citations and a log of search queries to a health information website.

Since the number of terms/concepts in these sources is quite large, studying co-occurrence data at the concept-level is too complex and may not be useful. Hence, we use a mechanism to group the concepts into broader categories and analyze co-occurrence information at the group level. This is done in two phases. First, we map the concepts to semantic types defined by Unified Medical Language System's (UMLS®) Semantic Network. UMLS' semantic network defines a set of subject categories, or semantic types that can be used to consistently categorize the more than million concepts defined by UMLS Metathesaurus® [11]. This aggregates all concepts in the three sources to a more manageable 135 categories.

Second, we partition the semantic types into 15 semantic groups using a partitioning scheme proposed by Bodenreider and McCray [12]. This scheme adequately adheres to partitioning principles such as semantic validity (the groups must be semantically coherent), completeness (the groups must cover the whole concept domain) and exclusivity (a concept must belong to a single group). Table 1 has a partial list of semantic groups and examples of semantic types in each group[1]. We refer to the semantic group of the semantic type to which a concept maps as the concept's semantic group

[1] A complete listing is available elsewhere [12]

and the exclusivity principle of partitioning ensures that a concept's semantic group is unique.

By analyzing this co-occurrence information at the semantic group level, we try to understand: a) the differences in prevalence of semantic groups in the three sources; b) the difference in mental model of a health consumer from that of a professional; c) the difference in professional-generated co-occurrence with change in the context of communication.

Table 1 - Common semantic groups and associated semantic types

Semantic Group	Semantic Type (TUI[2])
Concepts & Ideas	Quantitative Concept (T081); Functional Concept (T169); Qualitative Concept (T080); Temporal Concept (T079)
Disorder	Disease or Syndrome (T047); Finding (T033); Sign or Symptom (T184); Injury or Poisoning (T037)
Anatomy	Body part, Organ or Organ Component (T023); Body Location or region (T029); Body Space or Junction (T030)
Chemicals/ Drugs	Organic Chemical (T109); Clinical Drug (T200); Pharmacologic Substance (T121); Amino acid, peptide or protein (T116)
Procedures	Therapeutic or Preventive Procedure (T061); Laboratory Procedure (T059)

Materials and Methods

Data Collection

As mentioned in the previous section, we use concept-level co-occurrence information from two sources of professional-generated health content and one source of consumer-generated health content. The sources and definitions of co-occurrence in these sources are described below. Using these concept-level co-occurrence frequencies, we compute co-occurrence frequencies between semantic groups.

Research Patient Data Repository (RPDR)

This is a centralized repository of physician generated electronic medical reports such as discharge summaries and outpatient notes from several clinics and hospitals in the Partners HealthCare system [13]. Though the number of reports available through RPDR is quite large we used a subset of 5500 discharge summaries and mapped the reports' text to UMLS concepts[3]. The co-occurrence frequency of a pair of concepts is defined as the number of reports in which the two concepts occur in the same section of the report.

UMLS Co-occurrence (MRCOC)

The *mrcoc* table is distributed as part of the UMLS Metathesaurus and contains the co-occurrence frequencies of keywords in MEDLINE citations [15]. As MEDLINE is considered to be a comprehensive source of publications in biomedical journals, *mrcoc* is a good resource for studying concept co-occurrence in professional-generated health content.

MedlinePlus Query Log (QLOG):

MedlinePlus is an open-access website with an extensive collection of health information from the National Library of Medicine, National Institute of Health and other US government agencies targeting lay health information seekers [16]. Users can search the site for health topics of interest. We used an anonymized log of user queries to MedlinePlus as an example of consumer-generated co-occurrence. A pair of concepts is considered to have co-occurred if they (or the terms that map to these concepts) are queried in the same user session (from the same IP address within five minutes of each other)

Calculating co-occurrence of Semantic Groups

In the above sources, co-occurrence frequency is defined only at the concept-level and the frequency at the semantic-type and semantic-group level needs to be computed.

In a given source s, if the co-occurrence frequency of concepts c_i and c_j is $f_s(c_i, c_j)$, the co-occurrence frequency at the semantic-*type* level is defined as $f_s(t_p, t_q) = \sum_{i,j} f_s(c_i, c_j)$ where c_i is of semantic type t_p and c_j is of semantic type t_q. Similarly, the co-occurrence frequency at the semantic-*group* level can be defined as $f_s(g_m, g_n) = \sum_{p,q} f_s(t_p, t_q)$ where t_p belongs to semantic group g_m and t_q belongs to semantic group g_n.

Note that,

- $f_s(X, Y) = f_s(Y, X)$ at the concept-level and hence also at the semantic-type and semantic-group level;
- $f_s(c_i, c_j)$ is undefined if $i = j$, but $f_s(t_p, t_q)$ when $p = q$ and $f_s(g_m, g_n)$ when $m = n$ are both valid and well-defined;
- a concept can be mapped to more than one semantic type and in such cases needs to be considered in calculation of $f_s(t_p, t_q)$ for all t_x to which it maps. A semantic type, however, will belong to exactly one semantic group.

Results

Using the methods described above, we found 1.1 million co-occurrences in QLOG and a comparable number of co-occurrences - 0.8 million - in RPDR. MRCOC has larger number of unique concepts and hence a larger number of co-occurrences (11 million).

Figure 1 shows the co-occurrence distribution of the semantic groups in each source. For example, a value of 13% for the DISO semantic group in RPDR indicates that 13% of all co-occurrences defined in RPDR involved at least one concept of the semantic group DISO. As can be seen in the figure, in all the three sources, the semantic groups Anatomy, Chemical & Drugs, Concepts & Ideas, Disorders and Procedures together

[2] UMLS defined unique identifier for semantic type

[3] The text-to-concept mapping is done using the Health Information Text Extraction (HITEx) [14] system - an open source natural language processing tool.

account for about 85% of all co-occurrences. However in RPDR, *Concepts* is the dominating semantic group (33%) while in MRCOC *Chemicals and Drugs* contributes 43% of all co-occurrences. In contrast to these two professional-generated sources, the consumer-created QLOG data shows higher number of co-occurrences in concepts of semantic group *Disorders (32%)*.

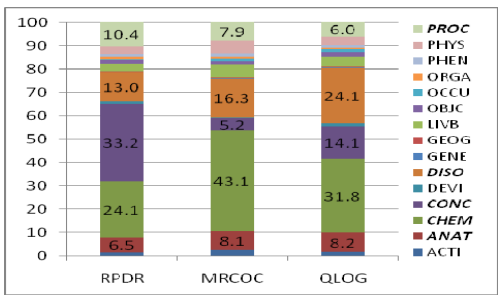

Figure 1- Co-occurrence distribution in each source

It is more interesting to look at co-occurrence between pairs of semantic groups in each of these sources. Figure 2 shows a graph[4] depicting the co-occurrences defined in QLOG. The nodes of the graph represent the semantic groups while the edges show the strength of co-occurrence between the semantic groups. The color and thickness of the edges is proportional to the strength of the co-occurrence between the semantic groups. Additionally, to reduce clutter only the top 20% of the edges have been shown in the graph and loop-edges (source = target) were ignored. Figure 3 and Figure 4 show the graphs for RPDR and MRCOC respectively.

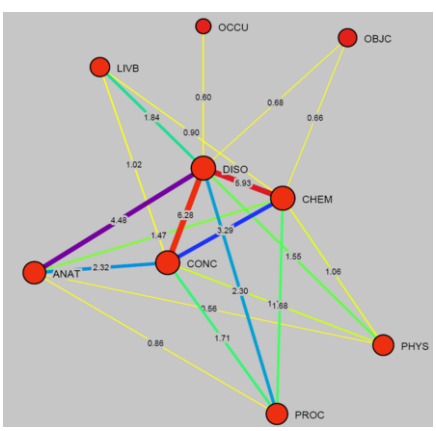

Figure 2 - Co-occurrence in QLOG

In Figure 2 it can be observed that the top three edges are DISO-CONC (6.28%), DISO-CHEM (5.93%) and DISO-ANAT (4.48%) and all three involve concepts of DISO. In RPDR (Figure 3) similar dominance is noted for concepts of CONC with the top three edges being CONC-CHEM (14.4%), CONC-DISO (7.94%) and CONC-PROC (6.12%), while the other professional-generated source MRCOC (Figure 4) has a greater representation from concepts of type Chemicals & Drugs - CHEM-DISO (9.89%), CHEM-ANAT (6.50%) and CHEM-PHYS (4.05%).

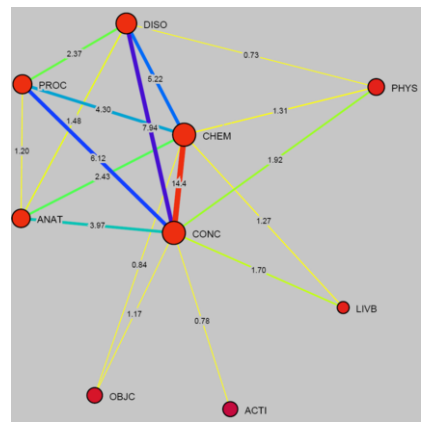

Figure 3 - Co-occurrence in RPDR

We believe that these graphs represent a difference in a consumers' interest in the medical domain and the professionals' interest. The consumers appear to be more inclined to learn about a disease condition and concepts of other semantic groups are explored in their relation to the DISO concept. The other two sources have different focus and the information retrieved from these sources may not be of equal interest to the consumers.

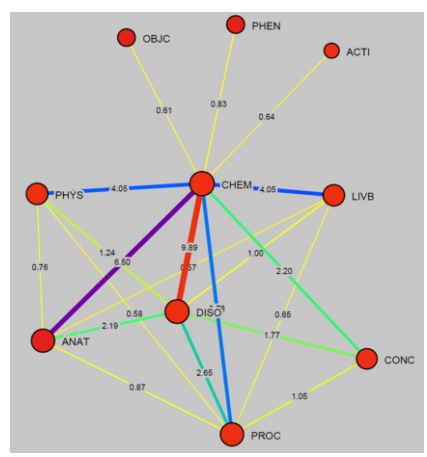

Figure 4 - Co-occurrence in MRCOC

[4] The graphs are generated using visualization software, Himmeli (v3.0.1), provided by the Folkhälsan Research Center at University of Helsinki, Finland.

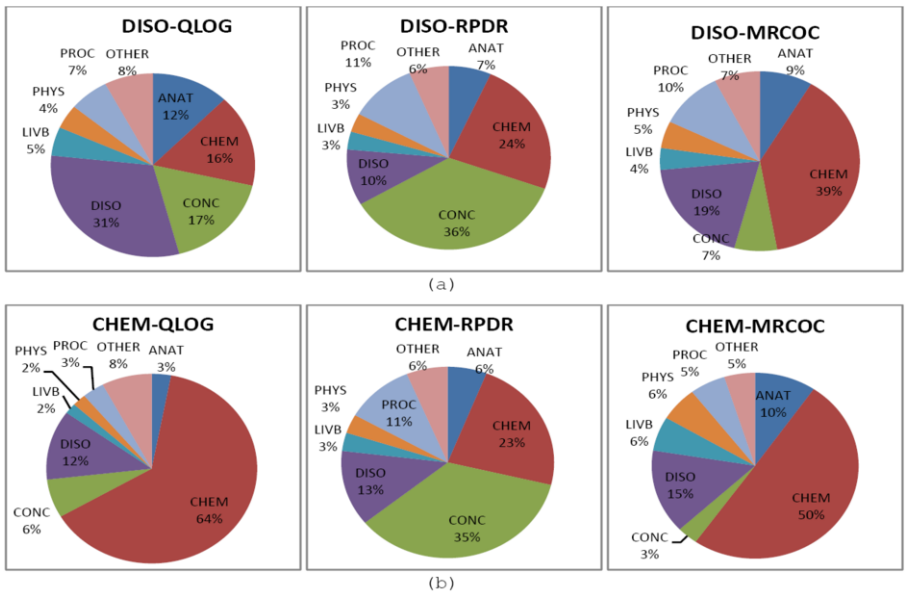

Figure 5 – Charts representing co-occurrence distribution for (a) DISO and (b) CHEM semantic groups in the three sources.

We analyzed the co-occurrence of DISO and CHEM semantic groups from the three sources in greater detail (Figure 5) as we believe they are of higher interest to consumers. The pie charts show the distribution of concepts among the semantic groups when one of the concepts of the co-occurring concept pair belongs to DISO (Figure 5(a)) and CHEM (Figure 5(b)).

Semantic group DISO

Figure 5(a) shows the co-occurrence of concepts of DISO semantic group with each other and concepts of other semantic groups in all three sources. For instance, the 12% ANAT in DISO-QLOG implies that in 12% of all co-occurrences defined in QLOG involving a concept of group DISO, the other concept is of type ANAT. Low frequency groups have been aggregated under OTHER.

DISO in QLOG shows a high intra-group co-occurrence (31%) which probably hints at scenarios where consumers query multiple symptoms they are experiencing or a disease name and follow it up with symptoms that are related to the condition. In neither RPDR nor MRCOC, DISO exhibits similar self co-occurrence. A related observation is the higher co-occurrence of ANAT with DISO concepts in QLOG (patients querying for body parts which are affected or in which the symptoms are manifested).

MRCOC shows a high co-occurrence between DISO and CHEM which may be indicative that in this source the disease terms co-occur with medications that are commonly prescribed to treat the disease or the results of experimental studies (since these are biomedical citations) on the efficacy of a chemical substance in the treatment of a condition. In RPDR, the high co-occurrence between DISO-CONC is consistent with the overall high prevalence of CONC in this source (as seen in Figure 1).

Semantic group CHEM

Figure 5(b) shows the corresponding pie charts for semantic group CHEM. Compared to DISO, CHEM shows a higher degree of intra-group co-occurrence in all three sources.

In QLOG, 64% of the CHEM concepts co-occur with other concepts of the same semantic group. This can mean that when consumers query for a medication or chemical substance they do so in the context of another term of similar type (substitutes or generic alternatives, for instance) or to understand its' chemical composition. The most significant inter-group co-occurrence is with concepts of group DISO (either as prescribed for or as a potential complication of). Similar distribution is observed in MRCOC except for a higher co-occurrence with concepts of ANAT.

CHEM in RPDR, on the other hand exhibits a much higher co-occurrence with concepts of CONC (35%) compared to QLOG (6%) and MRCOC (3%). The self-co-occurrence is also significantly lower.

Discussion

We found significant differences in the prevalence of the semantic groups and the strength of co-occurrences between semantic groups in the three sources. The co-occurrence defined by consumers differs from that defined by professionals. The two professional sources have noticeable differences with each other as well.

We believe these differences are a reflection of the mental models of the content creators in specific communication contexts: QLOG is authored by consumers in the context of information seeking, RPDR is authored by clinicians in the context of documenting patient care, and MRCOC is authored

by researchers (clinical as well as basic science) in the context of describing research studies findings.

In a way, the differences we have found are to be expected. Nevertheless, they have direct implications in consumer health informatics. Professional medical records, for instance, are the content source of many personal health record applications. The differences observed between QLOG and RPDR suggest that the content of medical records in its current form may not sufficiently satisfy patient information needs and has to be re-organized to facilitate information retrieval and understanding by patients.

For example, we have identified that consumers querying signs, symptoms or disorders tend to be very interested in associated signs, symptoms or disorders. While it is fairly easy to find diagnoses in professional or personal medical records, the relations between diagnoses and signs and symptoms are often not explained – this is partially reflected in the relatively low self co-occurrence in the DISO group. For the lay consumers to comprehend the content in their medical records, personal health record applications need to consider ways to help consumers connect the diagnoses with related signs and symptoms.

Similarly, consumers who queried for medications appeared to be very interested in other medications. In this regard, medical records are quite different – the self co-occurrence in the CHEM group in RPDR is 23% while it is 64% in QLOG. On the other hand, CHEM in MRCOC showed fairly high self co-occurrence (50%) compared to RPDR, suggesting that medications are discussed far more frequently in the context of other medications in biomedical literature than in medical records. While we would not expect an average consumer to use medical journals as the primary information source, the information in medical records may not be sufficient either.

We recognize that our analysis just scratched the surface in terms of understanding the layperson's and professional mental models. However, we hope the differences revealed by this study will help draw the consumer health informatics researchers' and developers' attention to the issue.

We also realize that our study can benefit from additional types of consumer generated content and we are looking into using data from online patient fora and from health-oriented social networking sites like PatientsLikeMe [17].

Acknowledgements

This work is supported by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK 075837) and NIH (R01 LM07222). We thank MedlinePlus for sharing their log data.

References

[1] Fox S, Jones S. The Social Life of Health Information. Pew Internet & American Life Project. Accessed June 11, 2009;Available from: http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx

[2] Nielsen-Bohlman L, Panzer AM, Kindig DA. Health Literacy: A Prescription to End Confusion. Washington, DC: National Academy Press 2004.

[3] Graham L, Tse T, Keselman A. Exploring user navigation during online health information seeking. AMIA Annual Symp Proc. 2006: 299-303

[4] Keselman A, Browne AC, Kaufman DR. Consumer health information seeking as hypothesis testing. JAMIA. 2008 Jul-Aug; 15(4);484-495

[5] Zeng QT, Crowell J, Plovnick RM, Kim E, Ngo L, Dibble E. Assisting consumer health information retrieval with query recommendations. JAMIA. 2006; 13(1):80-90.

[6] Zeng QT, Kogan S, Plovnick RM, Crowell J, Lacroix EM, Greenes RA. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. Int K Med Inform. 2004 Feb; 73(1):45-55.

[7] Slaughter L, Keselman A, Kushniruk A, Patel VL. A framework for capturing the interactions between laypersons' understanding of disease, information gathering behaviors, and actions taken during an epidemic. J Biomed Inform. 2005 Aug; 38(4):298:313

[8] Rotegrad AK, Slaughter L, Ruland CM. Mapping nurses' natural language to oncology patient's symptom expressions. Stud health Techn. Inform. 2006; 122:987-8.

[9] Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study J Med Internet Res 2007;9(1):e5

[10] Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC, Goryachev S, Ngo L Term Identification Methods for Consumer Health Vocabulary Development J Med Internet Res 2007;9(1):e4

[11] Available from: http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

[12] Bodenreider O, McCray AT Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003;36(6):414-432

[13] Available from: http://rc.partners.org/rpdr

[14] Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. 2006; 6(1):30.

[15] Available from: http://www.nlm.nih.gov/databases/databases_medline.html

[16] Available from: http://medlineplus.gov/

[17] Available from: http://www.patientslikeme.com/

Address for Correspondence

Qing Zeng-Treitler, PhD
Department of Biomedical Informatics,
University of Utah
26 S 2000 E, HSEB 5700,
Salt Lake City, UT. 84112.
Email: q.t.zeng@utah.edu