Determining Prominent Subdomains in Medicine

Powell J. Bernhardt,[a] Susanne M. Humphrey,[b] Thomas C. Rindflesch[b]
[a]*Temple University, Philadelphia, Pennsylvania*
[b]*National Library of Medicine, Bethesda, Maryland*

We discuss an automated method for identifying prominent subdomains in medicine. The motivation is to enhance the results of natural language processing by focusing on sublanguages associated with medical specialties concerned with prevalent disorders. At the core of our approach is a statistical system for topical categorization of medical text. A method based on epidemiological evidence is compared to another that considers frequency of occurrence of Medline citations. We suggest the isolation of UMLS terminology peculiar to individual medical specialties as a way of enhancing natural language processing systems in the biomedical domain.

INTRODUCTION

As quality assurance and risk management continue to be major issues in the delivery of safe and effective health care, evidence-based medicine is an appropriate strategy for implementation, supported by automatic access to the biomedical literature. Natural language processing offers methods of extracting useful information from biomedical text for a range of clinical and research applications. Due to the complexity of natural language, such applications are often limited by text genre, primarily patient records [1,2] or the medical literature [3,4].

Natural language processing systems are also characterized by the domain in which they apply. In medicine, some are limited to the clinical area [5,6] and others to molecular biology [7,8]. Within the clinical domain, some phenomena could be addressed more effectively by further limiting processing to specific areas, such as cardiology for resolving the abbreviations in (1). (Also see [9,10]).

(1) **ICD** implantation in patients with **CAD**, unexplained syncope and inducible **VF**

Harris introduced the notion of a sublanguage [11], that is, a subset of language structures and phenomena used in a particular domain. The theory of sublanguage has been discussed as a vehicle for improving the quality of natural language processing in medicine, particularly for clinical medicine and genomics [12]. Methods of exploiting the notion of sublanguage to improve results in areas other than medicine have also been proposed [13].

Because considerable effort must be expended in crafting a natural language processing system to accommodate a sublanguage, it is important, as a first step, to determine "prominent" subdomains in medicine, that is, areas that have large amounts of relevant text. In this paper, we propose an automated method for accomplishing this, by considering medical specialties as the basis for the subdomains, assuming that each specialty has its own sublanguage. (See [14].)

BACKGROUND

Medical Specialties

We first consider criteria for isolating prominent subdomains in medicine, paying attention to the appropriate level of granularity. Medicine itself is a domain (in contrast, say, to business or law), however, we seek a finer level of granularity. It might be possible to define sublanguages at the level of diseases, but we pursue the medical specialties as a useful level of granularity for focusing natural language processing systems.

The medical specialties in the United States and Western Europe have developed as they are for sociological as well as medical reasons [15,16]; however, they are categorized medically according to several criteria. Some apply to anatomic organ or body system. Patient population, pathological process, intervention, and the nature of the problem classify others [17]. Pediatrics is categorized by patient population, for example, whereas cardiology is classified by body system.

Classification Research

In order to manipulate the medical specialties for determining medical sublanguages we rely on classification research, using the National Library of Medicine's journal descriptors for characterizing text. These terms constitute a library classification used for organizing knowledge in documents. Satija [18] contrasts this with an actual knowledge classification; however, journal descriptors are also a special classification (for a specific area of knowledge), in contrast to a general, or universal, classification. In fact, many of them correspond to titles of subclasses in the Library of Congress Classification (itself a universal classification), specifically, CLASS R - MEDICINE and CLASS Q – SCIENCE [19].

Journal descriptors are a set of 127 MeSH indexing terms (for example, Cardiology, Pediatrics, Surgery, Emergency Medicine, and Brain) used by NLM to index journals *per se*. For example, the Journal of Pediatric Surgery is indexed by the journal descriptors Pediatrics and Surgery. Being discipline-based, the journal descriptor classification can be said to reflect certain epistemological views. Hjorland and Albrechtsen [20] state that a classification that scatters subjects by discipline, and thus human interests, is an expression of a philosophy of knowledge combining historicism (based on the development of knowledge producing communities, i.e., the division of scientific labor) and pragmatism (based on the development and state of knowledge). They discuss the Dewey Decimal Classification as an example.

It should be noted that journal descriptors reflect subject areas of journals. Thus, although expressed predominantly in "study of" (e.g., "ology") type terminology, they include some terms for organs, diseases, facilities, drugs, procedures, processes, etc. The names for the medical specialties are a subset of the journal descriptors, and in this study we limit processing to them. We use a method for exploiting journal descriptors called Journal Descriptor Indexing (JDI) [21,22].

Journal Descriptor Indexing

JDI is a fully automated indexing tool for documents in the biomedical domain. Topical categorization is based on the association between text in Medline citations and journal descriptors. In particular, the JDI system associates journal descriptors with words in titles and abstracts in a training set of Medline records. The version used in this research is a one-year training set of 435,300 records. Each record in the training set "inherits" the journal descriptors from the journal in that record. A word in the training set can be described by a list of journal descriptors ranked according to the number of co-occurrences between the word and the journal descriptors. Text as input to the JDI system can be indexed based on averaging the word-journal descriptor cooccurrences for the words in the text that are also in the training set, ranking the journal descriptors in decreasing order of these averages. For example, JDI associates the text (2) with medical specialties surgery and pediatrics.

(2) The charts of all children undergoing appendectomy between 1988 and 1998 were analyzed.

METHODS

A reasonable approach to determining prominent subdomains in medicine is to concentrate on those medical specialties concerned with prevalent disorders. However, a particular disorder may pertain to more than one specialty. For example, a journal article about the effectiveness of a new intervention for acute myocardial infarction may be of interest to several medical specialties. In order to accommodate the interaction of disorders and specialties, we first determine prevalent disorders in the United States by relying on epidemiological reports as well as frequency of occurrence of MeSH terms in Medline citations. We then use JDI to associate the most common disorders with the specialties involved.

Determining Prevalent Diseases

We obtained prominent disorders based on the most frequent primary diagnoses groups and causes of mortality from the Centers for Disease Control and Prevention as reported in the National Ambulatory Medical Care Survey [23]. For this project, we omitted the primary diagnoses group "general medical examination" because it is not a disorder. We also obtained the ten most common causes of death from a report by the same agency [24]. Finally, in order to tie epidemiological information with the medical literature, we used MetaMap [25] to map diagnoses and causes of death to MeSH terms, which may be preferred names or their synonyms. The most common causes of death are given in (3), excluding non-diseases accidents, suicide, and homicide. The most frequent diagnoses are given in (4), excluding child health services, pregnancy, and physical examination. The MeSH equivalents are given (indented, below the CDC terms) in both lists.

(3) Diseases of heart
    Heart Diseases
  Malignant neoplasms
    Cancer
  Cerebrovascular diseases
    Cerebrovascular Disorders
  Chronic obstructive pulmonary diseases
    Lung Diseases, Obstructive
  Diabetes mellitus
    Diabetes Mellitus
  Pneumonia and influenza
    Pneumonia
  Chronic liver disease and cirrhosis
    Diseases, Liver
  Human Immunodeficiency Virus
    HIV

(4) Essential hypertension
    Hypertension

Arthropathies and related disorders
Joint Diseases
Acute upper respiratory infections, excluding pharyngitis
Respiratory Tract Infections
Diabetes mellitus
Diabetes Mellitus
Spinal disorders
Spinal Diseases
Rheumatism, excluding back
Rheumatic Diseases
Malignant neoplasms
Cancer
Heart disease, excluding ischemic
Heart Diseases

In further manipulating these lists we used the equivalent MeSH terms and combined the leading causes of death with the most frequent diagnoses, eliminating duplicates (Heart Diseases, Cancer, Diabetes Mellitus); the remaining twelve terms are given as (5).

(5) Heart Diseases
Cancer
Cerebrovascular Disorders
Lung Diseases, Obstructive
Diabetes Mellitus
Pneumonia
Diseases, Liver
HIV
Hypertension
Joint Diseases
Spinal Diseases
Rheumatic Diseases

We then determined the frequency of the MeSH terms in Medline citations, sorted them according to frequency, and retained the ten most frequent (6). In combining information from epidemiology and the medical literature on these phenomena, we provide a more accurate representation of actual prominence.

(6) Cancer (113,662)
Hypertension (110,319)
Diabetes Mellitus (54,059)
Liver Diseases (38,913)
Cerebrovascular Disorders (35,588)
Heart Diseases (31,385)
Pneumonia (21,144)
Respiratory Tract Infections (18,425)
Lung Diseases, Obstructive (16,971)
Joint Diseases (14,172)

JDI was used to compute the interaction of disorders with medical specialties. For each prevalent disorder in (6), we retained the top two journal descriptors (limited to the specialties) returned by JDI, for example, Cardiology and Pulmonary Disease (Specialty) for Heart Disease. Complete results are given in Table 1.

Prevalent Diseases	**Top 2 JDs limited to Medical Specialties**
Cancer	Medical Oncology, Urology
Hypertension	Nephrology, Cardiology
Diabetes Mellitus	Endocrinology, Nephrology
Liver Disease	Gastroenterology, Toxicology
Cerebrovascular Disorders	Neurosurgery, Neurology
Heart Diseases	Cardiology, Pulmonary Disease (Specialty)
Pneumonia	Pulmonary Disease (Specialty), Communicable Diseases
Respiratory Tract Infections	Communicable Diseases, Pulmonary Disease (Specialty)
Lung Diseases, Obstructive	Pulmonary Disease (Specialty), Medical Oncology
Joint Diseases	Orthopedics, Rheumatology

Table 1. Prevalent Diseases with Journal Descriptors

Evaluation
We conducted an assessment of the method just described by comparing it to one that relies exclusively on the medical literature. A PubMed search using the MeSH subheading “therapy” and limited to six months of Medline retrieved 23,800 citations. We processed these with JDI and kept the single top journal descriptor (again limited to the medical specialties). After sorting and counting, we retained the twenty most frequent.

RESULTS

Based on JDI processing of prevalent disorders (Table 1), the most frequently associated medical specialties are listed as (7).

(7) Pulmonary Disease (Specialty) (4)
Cardiology (2)
Communicable Diseases (2)
Medical Oncology (2)
Nephrology (2)
Endocrinology, Gastroenterology, Neurology, Neurosurgery, Orthopedics , Rheumatology, Toxicology, Urology (1)

The twenty most frequently occurring specialties computed by examining Medline citations are given split into groups of ten in (8), including the number of relevant citations as determined by JDI.

(8) 1344 Orthopedics
1293 Medical Oncology
1209 Cardiology
1037 Gastroenterology
999 Ophthalmology
921 Urology
918 Psychiatry
738 Pulmonary Disease (Specialty)
588 Otolaryngology
578 Endocrinology

570 Hematology
533 Anesthesiology
507 Dermatology
488 Nephrology
484 Neurosurgery
478 Communicable Diseases
383 Surgery
340 Neurology
302 Rheumatology
282 Obstetrics

DISCUSSION

JDI addresses the interaction of diseases and the medical specialties concerned with them, as indicated in Table 1. One specialty may deal with more than one class of disorder, as for example the association of Pulmonary Disease (Specialty) with Heart Diseases as well as Pneumonia, Respiratory Tract Infections, and Lung Disease, Obstructive. On the other hand, a single disease may be relevant to more than one specialty; for example JDI assigned both Nephrology and Cardiology to Hypertension.

There is considerable agreement in the two informatics methods used to determine the most prominent subdomains in medicine. Of the thirteen specialties indicated as being concerned with prevalent diseases by the first method in (7), seven also occur in the ten most frequent based on number of citations in Medline in (8): Pulmonary Disease (Specialty), Cardiology, Medical Oncology, Endocrinology, Gastroenterology, Orthopedics, and Urology. An additional five specialties identified as prominent by association with prevalent disorders fall within the next ten most frequent in Medline: Communicable Diseases, Nephrology, Neurology, Neurosurgery, Rheumatology. According to these computations, only Toxicology is associated with a prevalent disorder but is not frequently discussed in the medical literature.

We plan to exploit the results of this study by devising methods for isolating UMLS terminology pertinent to the most prominent medical specialties as determined by our method, beginning with Pulmonary Disease (Specialty), Cardiology, Medical Oncology, Nephrology, and Endocrinology. We intend to take advantage of hierarchical structure in the Metathesaurus as well as semantic types related to disorders, anatomy, physiology, and procedures. (Also see [26]). It is then possible to use JDI to identify the specialty of text being processed. Focusing on the relevant sublanguage (in particular, terminology) can then enhance the results of natural language processing, thereby providing useful information in support of evidence-based practice.

CONCLUSION

This study was motivated by the need to enhance effectiveness of natural language processing in the medical domain. Guided by the principles of sublanguage theory we developed a method of identifying prominent subdomains in medicine that combines information from epidemiology and the medical literature and relies on JDI, an automated technique for topical categorization of biomedical text. We discuss prospects for exploiting the results of this project to help craft natural language processing systems by focusing on terminology peculiar to individual medical specialties.

Acknowledgements

We thank Halil Kilicoglu for his contributions to this project and are grateful to Danesh Mazloomdoost and Marcelo Fiszman for comments on an earlier version of the paper. The first author was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

1. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7:593-604.
2. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automatic encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11:392-402.

3. Mendonça EA, Johnson SB, Seol Y, Cimino JJ. Analyzing the Semantics of Patient Data to Rank Records of Literature Retrieval. Proc ACL Workshop on Natural Language Processing in the Biomedical Domain 2002;69-76.
4. Rindflesch TC, Fiszman M. The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text. J Biomed Inform 2003;36(6):462-77.
5. Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp 2000;235-9.
6. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. MEDINFO 2004;487-91.
7. Friedman C, Dra P, Yu J, Krauthammer M, Rzhetsky A. GENIES: a natural language processing system for the extraction of molecular pathways from journal article. Bioinformatics 2001;17(suppl 1):s74-82.
8. Libbus B., Kilicoglu H., Rindflesch TC, Mork, J G, Aronson AR. Using Natural Language Processing, Locus Link, and the Gene Ontology to Compare OMIM to MEDLINE. Proc HLT-NAACL Workshop on Linking the Biological Literature, Ontologies and Databases: Tools for Users 2004;69-76.
9. Liu H, Aronson AR, Friedman C. A study of abbreviations in MEDLINE abstracts. Proc AMIA Symp 2002;464-9.
10. Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in electronic articles. J AM Med Inform Assoc 2002;9(3):262-72.
11. Harris Z. A theory of language and information: a mathematical approach. Oxford: Clarendon Press; 1991.
12. Friedman C, Kra P, Rzhetsky. Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 2002;35:222-35.
13. Grishman R, Kittredge R. Analyzing language in restricted domains: sublanguage description and processing. Hillsdale, NJ: Lawrence Erlbaum; 1986.
14. Starren J, Johnson SM. Notations for high efficiency data presentation in mammography. Proc AMIA Annu Fall Symp. 1996;:557-61.
15. Starr, P. The social transformation of American medicine: the rise of a sovereign profession and the making of a vast industry. New York: Basic Books; 1982.
16. Stevens, R. American medicine and the public interest, updated ed. with new introduction. Berkeley: University of California Press; 1998.
17. www.abms.org/approved.asp
18. Satija MP. Library classification: an essay in terminology. Knowledge Organization 2000;97(4):221-9.
19. http://www.loc.gov/catdir/cpso/lcco/lcco.html
20. Hjorland P, Albrechtsen H. An analysis of some trends in classification research. Knowledge Organization 1999;26(3):131-9.
21. Humphrey SM. Automatic indexing of documents from journal descriptors: a preliminary investigation. J Am Soc Inf Sci 1999;50:661-74.
22. Humphrey SM, Rindflesch TC, Aronson AR. Automatic indexing by discipline and high-level categories: methodologies and potential applications. Proc 11th ASIST SIG/CR Classification Research Workshop 2000;103-16.
23. National Ambulatory Medical Care Survey. Number and percent distribution of office visits with corresponding standard errors, by selected primary diagnosis groups and patient's sex: United States, 2001 Source: U.S. Department of Health & Human Services, Public Health Service, Centers for Disease Control & Prevention, National Center for Health Statistics, 2001 data.
24. Centers for Disease Control and Prevention's Health, United States, 2003 With Chartbook on Trends in the Health of Americans.
25. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp 2001;17-21.
26. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proc AMIA Symp 2001;189-93.