MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Application of a Medical Text Indexer to an Online Dermatology Atlas

GR Kim[a], AR Aronson[b], JG Mork[b], BA Cohen[c], CU Lehmann[a]

[a] *Division of Health Sciences Informatics, Johns Hopkins University, Baltimore, MD,*
[b] *National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, MD,*
[c] *Division of Pediatric Dermatology, Johns Hopkins University, Baltimore, MD*

Abstract

Clinical dermatology cases are presented as images and semi-structured text describing skin lesions and their relationships to disease. Metadata assignment to such cases is hampered by lack of a standardized dermatology vocabulary and facilitated methods for indexing legacy collections. In this pilot study descriptive clinical text from Dermatlas, a Web-based repository of dermatology cases, was indexed to Medical Subject Heading (MeSH®) terms using the National Library of Medicine's Medical Text Indexer (MTI). The MTI is an automated text processing system that derives ranked lists of MeSH terms to describe the content of medical journal citations using knowledge from the Unified Medical Language System® (UMLS®) and from MEDLINE®. For a representative, random sample of 50 Dermatlas cases, the MTI frequently derived MeSH indexing terms that matched expert-assigned terms for Diagnoses (88%), Lesion Types (72%), and Patient Characteristics (Gender and Age Groups, 62% and 84% respectively). This pilot demonstrates the potential for extending the MTI to automate indexing of clinical case presentations and for using MeSH to describe aspects of clinical dermatology.

Keywords:

Medical Informatics, Dermatology, Natural Language Processing, Controlled Vocabulary

Introduction

Network technology (NT) enables clinicians to share clinical experience and medical knowledge with remote colleagues for patient care, research and teaching. With multimedia presentations and distributed environments such as the World Wide Web, individual clinicians can make traditional forms such as case presentations reproducibly available to wide audiences. By combining NT with database technology, groups of clinicians, teachers and researchers can archive case presentations into sharable repositories of reusable clinical teaching materials for widespread teaching, reference and retrieval.

As such electronic repositories grow, their usability becomes increasingly dependent on indexing or assignment of metadata from controlled vocabularies or lexicons [1] to describe the content of individual documents. The power of the indexing depends on the completeness and granularity of the controlled vocabulary to describe the content of documents adequately and on the consistency of the assignment of metadata to describe similar documents. The power of assigned metadata can be extended if they can associate content from multiple information sources, as in a MEDLINE® InfoButton [2], where entries from an electronic medical record (EMR) laboratory result panel map to Medical Subject Heading (MeSH®) terms to create queries that link related journal abstracts from the MEDLINE citation database [3] to the medical record.

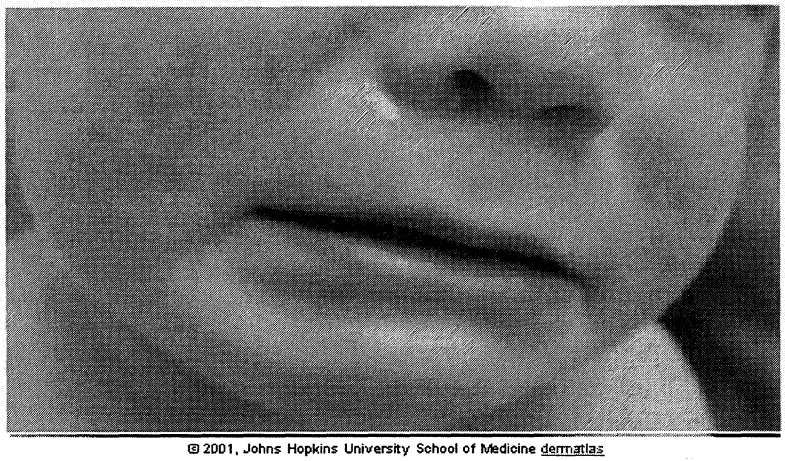
As collections continue to grow, by addition of new documents and/or incorporation of legacy archives, manual assignment of metadata to every document becomes less feasible and more expensive, particularly in technical domains such as medicine. These considerations have led to development and implementation of automated and semi-automated methods of indexing of documents such as medical journal citations [4].

This pilot study was performed to investigate two questions: 1) Can a system designed to automate or semi-automate the indexing of medical journal citations be extended to perform the same function for a library of short clinical text case presentations? and 2) What are strengths and weaknesses of using Medical Subject Headings (MeSH), a vocabulary designed for indexing medical journal articles, to describe diagnoses and other entities from clinical dermatology?

Materials and Methods

The indexing system used is the Medical Text Indexer (MTI) [5], from the National Library of Medicine (NLM). The MTI is the central software application in NLM's Indexing Initiative (II) [4][6] to investigate the feasibility of substituting automated or semi-automated methods for current domain expert-based indexing practices. The MTI applies alternative methods, based on semantic techniques derived from the Unified Medical Language System® (UMLS®) and statistical information from the MEDLINE database of citations, to compute ordered lists of indexing terms that describe the text content of titles and abstracts of medical journal citations.

The controlled vocabulary explored, MeSH, also from the National Library of Medicine, is used to index citations to full-text articles listed in the MEDLINE database. "MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts" [7]. At this time, there is no universally accepted formal terminology to describe clinical dermatology, although one is in development [1].



© 2001, Johns Hopkins University School of Medicine dermatlas

Image Name:	exanthem_1_010811	File Type:	jpg
Diagnosis:	FIFTH DISEASE / ERYTHEMA INFECTIOSUM VIRAL INFECTIONS, EXANTHEM	Category:	infections and infestations
Body Site:	cheek / face	Age:	21 months
Gender:	Male	Image Year:	2001
Contributor:	Brian Corden, MD	First Published:	July 30, 2001
Description:	diffuse red papular eruption with confluence on the cheeks		
Comments:	A 21 month old boy ran a high fever with irritability and congestion for 3 days. On the fourth day he developed a widespread morbilliform eruption that persisted for 6 weeks. Parvovirus B19 titers were positive for acute infection		
Related Links:	eMedicine - Erythema Infectiosum (Fifth Disease) Hardin Meta Directory: Fifth's Disease		

Figure 1 - Online Dermatlas entry

The clinical dermatology case repository used is Dermatlas [8] from the Johns Hopkins University School of Medicine. Released in December 2000, Dermatlas provides online access to high quality clinical and histologic images with text descriptions of findings and associated information related to over 3500 adult and pediatric dermatology cases. One Dermatlas feature links cases to external sources of related information (such as PubMed) via directed queries. Currently, keywords for such queries are arbitrarily assigned.

Each Dermatlas case contains an English text description of the clinical aspects of the case linked to a clinical image. Each case has fields for principal diagnosis, disease category and keywords (author-assigned) and text to describe clinical presentation, anatomic lesions and pathophysiology associated with the case. Additional fields are provided for other metadata if available (gender, age, pigmentation, morphology, anatomic location, pattern, organization, lesion color, date of photograph).

Fifty Dermatlas cases (Example: Figure 1, 2) were independently reviewed by 3 domain experts (GRK, CUL, BAC) who assigned MeSH terms through a consensus process [9] to a predetermined list of fields to describe each entry: Diagnosis, Disease Group, Lesion Type, Anatomic Location, Exposures, Diagnostic Procedures, Therapeutics, Gender and Age Group. If multiple MeSH terms were deemed appropriate for a field, all were entered.

The same cases were submitted to the MTI for indexing in the format shown in Figure 2. The MTI extracts noun phrases from text and maps them to UMLS concepts that in turn are mapped to MeSH terms via a restriction algorithm to produce initial recommendations. The MTI also adds ranked MeSH terms from MEDLINE citations that are similar to the case under consideration (PubMed Related Citations algorithm [10]) to the recommendations. The recommended terms are then filtered according to one of three selected filters to remove inappropriate recommendations. The low filtering option (Base) removes terms known to be unhelpful for indexing and provides a mixed list of good and bad recommendations with a fair number of good recommendations near the top of the list. The high filtering option (Tweak1) is the most exclusive and tends to give a small list of good recommendations but also filters out other good recommendations. The medium filtering option (Tweak2) is more lenient than Tweak1 and uses ten heuristics to balance the results by removing spurious and general terms when a more specific term is found ("general" and "specific" being determined by the MeSH tree hierarchy). Tweak2 provides a good-sized list with mostly correct recommendations [6][11][12]. Tweak2 was used in this pilot. The aggregate lists of MeSH terms assigned by the experts and the MTI were compared.

Results

A comparison of expert-assigned and MTI-assigned MeSH terms to describe the 50 Dermatlas entries is summarized in Table 1.

UI - 2100305770
TI - fifth disease, erythema infectiosum, viral infections, exanthem
AB - infections and infestations, erythema infectiosum, fifth disease, Parvovirus B19, diffuse red papular eruption with confluence on the cheeks, A 21 month old boy ran a high fever with irritability and congestion for 3 days. On the fourth day he developed a widespread morbilliform eruption that persisted for 6 weeks. Parvovirus B19 titers were positive for acute infection, 21.0 months, Year 2001, cheek, face

DermAtlasID	2100305770		
Diagnosis	Erythema Infectiosum		
MeSH Disease Class	Virus Diseases	Parvovirus B19, Human	Nasopharyngeal Diseases
Lesion Type	Exanthema	Fever	Irritable Mood
Anatomic Location	Face	Cheek	
Exposure	Virus Diseases		
Procedure	Antibodies, Viral		
Therapy			
Gender	Male		
AgeGroup	Infant		

2100305770|Erythema Infectiosum|C0085273|115164|MH|TI|MM;RC
2100305770|Exanthema|C0015230|14109|MH|TI;AB|MM
2100305770|Parvovirus B19, Human|C0085274|8504|MH||RC
2100305770|Herpesviridae Infections|C0019372|1110|MH|TI|MM
2100305770|Parvovirus|C0086776|700|MH|AB|MM

Figure 2 - Dermatlas entry (text portion), Expert-assignment and MTI-assignment of MeSH terms

In each case, notation was made if the MTI provided at least one match for the expert-assigned MeSH term, if it did not provide a matching term, if it provided a "reasonable" alternative (domain expert judgment) or if it provided a term that was wrong or contrary to the meaning or content of the entry (domain expert judgment).

In most cases, the expert indexers were able to find satisfactory MeSH terms to represent the content of the Dermatlas entries. In one principal diagnosis (Erythema Annulare Centrifigum), a satisfactory MeSH term to the proper granularity could not be found (MTI assigned "Erythema" as an indexing term to the entry). There was a high frequency of agreement between MeSH terms that experts assigned as "Diagnoses" and the terms that the MTI ranked as first or second to describe an entry (40 out of 44 cases).

In two cases, the MTI assigned MeSH terms that were not in the expert-assigned standard for "Diagnoses" ("Letterer-Siwe Disease" as an additional term for "Histiocytosis X", "Alopecia" for "Trichotillomania"), but which on review were considered very appropriate and important terms. In three cases, the MTI derived MeSH terms that were less specific than those assigned by experts, but which had a similar meaning ("Tinea" instead of "Tinea Corporis"). In one case, the MTI provided a MeSH term plus a Subheading ("Salivary Glands, Minor"[MeSH] + "injuries"[Subheading]) instead of the terms that the experts assigned ("Salivary Glands"[MeSH] + "Trauma"[MeSH]), and these were judged as equivalent. The MTI also frequently assigned MeSH terms that were assumed by the experts to be true for all cases within Dermatlas ("Humans" and "Skin Diseases").

Discussion

The automation of indexing text documents and other semi-structured information has been explored in different medical domains such as imaging, electrocardiography, pathology, hospital discharges [13][14][15][16] and medical journal indexing [17]. Natural language processing (NLP) systems that map free text from different types of medical documents to controlled vocabularies have been effective in limited domains with extensibility to related domains and document types [15]. This pilot explores the extension of a system and vocabulary for indexing medical journal citations to the domain of dermatology case presentations.

The possibility of using the MTI was suggested by the general similarity in structure of Dermatlas entries and MEDLINE citations: a title line with a semi-structured text description or abstract written in technical medical language. Case presentations use clinical language in a manner similar to that of journal articles and medical records. The abundant case material from Dermatlas provided the substrate (and opportunity) for this and future exploration.

The lack of a recognized controlled vocabulary for clinical dermatology suggested the exploration of an existing vocabulary. This, combined with the wish to be able to link Dermatlas cases with relevant citations from MEDLINE and the availability of the MTI, made MeSH the most apparent choice for exploration. A national group of clinical and medical terminology experts is currently developing a controlled vocabulary for dermatology and the results of this pilot may help guide its development.

Table 1: Expert and MTI Assignment of MeSH Terms to Dermatlas Entries

Field	Human	%MTI	%No MTI	%Alt	#Wrong
Diagnosis	49	87.8	6.1	2	1
Disease Category	49	44.9	26.5	28.6	9
Lesion Type	50	72	18	10	4
Anatomic Location	48	39.6	35.4	22.9	1
Exposure	19	36.8	52.6	0	1
Procedure (Dx)	12	58.3	33.3	0	1
Therapy	8	50	50	0	1
Gender	42	61.9	38	N/A	0
Age Group	50	84	8	8	0

Human	Number of 50 cases where Expert assigned at least one MeSH term
%MTI	Percentage of cases where MTI matched Expert-assigned MeSH term
%No MTI	Percentage of cases where MTI did not match Expert-assigned MeSH term
%Alt	Percentage of cases where MTI provided a reasonable alternative (Review)
#Wrong	Number of cases where MTI provided an incorrect or contradictory term (Review)

This exploration of MeSH to describe clinical dermatology findings showed a number of interesting points. Overall, MeSH was fairly complete in its coverage of skin diseases within this set of case presentations with only one disease entity that could not be found. MeSH was not as complete (although it was fairly rich) in terms used to describe attributes of skin lesions. Although many terms describing specific lesions were available ("Blister"), some were embedded in other MeSH terms ("Vesicle" in "Skin Diseases, Vesiculobullous"), and some commonly used terms considered "standard" by dermatologists ("Lumps", "Bumps", "Plaques") were not found at all. Other topics specific to clinical dermatology as a specialty that appear to be under-represented in MeSH are dermatologic laboratory tests, epi-luminescence terminology, therapies and patch testing.

The MTI level of restriction used in this study (Tweak2) was the level that has produced the most useful lists of MeSH terms through experience for the reasons outlined. A cursory examination of lists produced by the alternative methods reinforced this.

The results of this pilot study are very promising and suggest directions for future exploration of automated indexing of dermatology case presentations (within Dermatlas and beyond). Another direction is providing data for the development of the Dermatology Lexicon Project [1]. Yet another is the use and evaluation of MTI-derived MeSH Terms to create concept-based queries to MEDLINE (to facilitate precise linkage and retrieval of associated information from online medical journals) from Dermatlas cases and other learning objects. Other areas of exploration include adjustment of MTI parameters to improve the quality of terms assigned to describe dermatology cases using MeSH or the clinical dermatology lexicon when it becomes available [1].

Acknowledgements

We gratefully acknowledge feedback and commentary by Susanne M. Humphrey from the National Library of Medicine and Harold P. Lehmann from the Johns Hopkins University Division of Health Science Informatics. Dr. George Kim was supported by National Library of Medicine Medical Informatics Training Grant T15 LM07452-01.

References

[1] Center for Future Health. DLP - Dermatology Lexicon Project. [Web site] 2002; URL: http://www.future-health.rochester.edu/dlp/ [Accessed 27 Aug 2003].

[2] Cimino JJ, Elhanan G and Zeng Q. Supporting

[3] Infobuttons with Terminological Knowledge. Proceedings of the AMIA Symposium, 1997; 4:528-32 (suppl). National Center for Biotechnology Information. Data bases. National Library of Medicine. [Web site] 2003; URL: http:// www.ncbi.nlm.nih.gov/Database/index.html [Accessed 27 Aug 2003].

[4] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch T, Wilbur WJ. The NLM indexing initiative. Proceedings of the AMIA Symposium 2000. 20:17-21 (suppl).

[5] National Library of Medicine. Medical Text Indexer (MTI). [Web site] 2003; URL: http://ii.nlm.nih.gov/mti.shtml [Accessed 27 Aug 2003].

[6] National Library of Medicine. NLM's Indexing Initiative, 2003. [Web site] URL: http://ii.nlm.nih.gov [Accessed 27 Aug 2003]

[7] National Library of Medicine. Medical Subject Headings. Medical Subject Headings. [Web site] 2003; URL: http:// www.nlm.nih.gov/mesh/meshhome.html [Accessed 28 Aug 2003]

[8] Lehmann CU, Cohen BA. Dermatlas.org – Dermatology Image Atlas. Johns Hopkins University [Web site] 2000; URL: http://www.dermatlas.org [Accessed 27 Aug 2003].

[9] Hripcsak G, Wilcox A. Reference Standards, Judges and Comparison Subjects: Roles for Experts in Evaluating System Performance. J Am Med Inform Assoc 2002; 9: 1-15.

[10]Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proceedings of the AMIA Symposium. 2001; pp 319-23.

[11]National Library of Medicine. Medical Text Indexer (MTI). [Web site] 2003; URL: http://ii.nlm.nih.gov/mti.shtml [Accessed 1 Sep 2003].

[12]National Library of Medicine. Medical Text Indexer (MTI) Processing Flow. [White paper] 2003; URL: http://ii.nlm.nih.gov/MTIMedical_Text_Indexer_Processing_Flow.pdf [Accessed 1 Sep 2003].

[13] Wagner MM. An automatic indexing method for medical documents. Proceedings of the Annual Symposium on Computer Applications in Medical Care, IEEE, 1991. pp. 1011-7.

[14]Krauthammer M, Hripcsak G. A knowledge model for the interpretation and visualization of NLP-parsed discharge summaries. Proceedings of the AMIA Fall Symposium, 2001. pp 339-43.

[15]Friedman C. A broad-coverage natural language processing system. Proceedings of the AMIA Fall Symposium, 2000. pp 270-4.

[16]Huang Y, Lowe HJ, Hersh W. A Pilot Study of Contextual UMLS Indexing to Improve the Precision of Concept-Based Representation in XML-Structured Clinical Radiology Reports. J Am Med Inform Assoc [serial online] 2003; Aug 4. URL: http://www.jamia.org/cgi/reprint/M1369v1.pdf [Accessed 1 Sep 2003].

[17]Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Fall Symposium, 2001. pp 17-19.

Address for correspondence:

George R. Kim, MD
Johns Hopkins University School of Medicine
Division of Health Sciences Informatics
2024 E. Monument Street 1-207
Baltimore, Maryland 2120