Medical Acronym Disambiguation Using Online Sources

Janet Rajan† Karen C. Davis† Pawel Matykiewicz‡ Wlodzislaw Duch∫ John Pestian‡

†Electrical and Computer Engineering
Department
University of Cincinnati
Cincinnati, OH USA 45221
janetr@gmail.com, karen.davis@uc.edu

‡Division of Biomedical Informatics
Department of Pediatrics
Cincinnati Children's Hospital Medical Center
Univ. of Cincinnati, Cincinnati, OH USA 45221
pawel.matykiewicz@gmail.com, john.pestian@cchmc.org

∫School of Computer Engineering, Nanyang Technological University, Singapore,
and Department of Informatics, Nicholaus Copernicus University,
Grudziadzka 5, 87-100 Torun, Poland

Abstract

Hospitals produce millions of patient records consisting of clinical annotations containing extensive usage of abbreviations. The data in these clinical annotations are an excellent source for bioinformatics research but the use of abbreviations can create ambiguity. The main objective of our research is to develop a software application that takes a medical acronym as input and accesses medical and pharmaceutical websites to retrieve information from articles containing the acronym along with a user selected full-form of the acronym. The retrieved information consists of article title, authors, publication date, article abstract, and Medical Subject Header (MeSH) details. The information is used by researchers in the Biomedical Informatics Division of the Cincinnati Children's Hospital Medical Center as part of a research effort for reducing the ambiguity created by the use of acronyms. Our contribution to this research effort is a framework for disambiguation that accesses online sources; we design and populate an internal database that can be used in future research efforts.

1. Introduction

Millions of text-based records are produced at hospitals every year. A patient's record is usually made up of more than 50 different types of clinical annotations that may include radiology reports, discharge summaries, and surgical and nursing notes. The data in these clinical annotations are an excellent source for bioinformatics research. Researchers at the Cincinnati Children's Hospital Medical Center (CCHMC), a large pediatric academic medical center, are developing software that will retrieve useful information from these clinical annotations without losing its meaning.

There is an extensive usage of abbreviations in clinical annotations as it makes it easier to write the notes. But the use of abbreviations can create ambiguity, and what the term stands for could vary from one usage to another. For example the term 'PCA' may stand for 'Passive Cutaneous Anaphylactic,' 'Percutaneous Carotid Angiography,' 'Patient Controlled Anesthesia' or 'Physicians Corporation of America.' This kind of ambiguity created by the use of abbreviations (or acronyms) can cause problems in the retrieval of data in a systematic manner for further research. We use the terms *acronym* and *abbreviation* interchangeably in this paper.

Our work is a part of a larger project at CCHMC to develop a tool to handle uncontrolled use of abbreviations. For this purpose, freely available search and extraction tools that help retrieve abbreviations with their full-forms are used. Then medical articles containing a selected full-form are retrieved and filtered to keep only those with Medical Subject Headers (MeSH) [1]. The MeSH headers will be used by bioinformatics researchers at CCHMC in further research to determine whether MeSH data can be used to narrow down the definition or full-form of an acronym.

MeSH is the National Library of Medicine's controlled vocabulary thesaurus. A controlled vocabulary keeps track of terms related to each other by maintaining a list of words that are used to tag units of information so that the tagged information may be easily retrieved by a search [2]. MeSH consists of sets of named descriptors in a hierarchical structure that permits searching at various levels of specificity. The MeSH thesaurus is used by NLM for indexing articles from nearly 5000 of the world's leading biomedical journals for the MEDLINE/Pubmed database [1].

Pubmed is the digital archive of life sciences journal literature in the National Library of Medicine (NLM) developed and maintained by the National Center for Biotechnology Information (NCBI) which is a part of the

U.S. National Institutes of Health (NIH). With Pubmed, NLM is able to provide access to electronic literature and hence ensure the durability and utility of the archive as technology changes over time [3].

Medilexicon has a very large online database of over 200,000 pharmaceutical and medical abbreviations. Their database is updated daily to include new acronyms and their full-forms. Their resources are free to use, allowing people to easily look up full-forms for acronyms from the fields of medicine, pharmacy, bio-technology, and healthcare [4].

Our software tool builds a dictionary of medical abbreviations using Medilexicon to select a full-form for an acronym and then links the full-form of the acronym to medical articles in which these full-forms appear along with the acronym. The medical articles are retrieved from the Pubmed database. These articles also have related MeSH headers listed within their structure. Of the retrieved articles where a specific full-form and acronym pair is found, we retain only those articles containing MeSH headings. Our system compiles data from various web-based sources into a database and this information is then accessible for systematic retrieval at a later time. The user is provided with a facility to retrieve information from the internal database and view it.

For example if we looked for the acronym 'PCA,' we would first retrieve all full-forms of the term and then retrieve all of the articles that contain the acronym along with one selected full-form, say 'Passive Cutaneous Anaphylactic.' Once these articles are retrieved we keep information regarding only those that have MeSH headings. Once we have collected the MeSH headers from all the articles containing a given full-form of an abbreviation, we use it to calculate the number of times a particular acronym/full-form pair appears along with a specific MeSH header. This ratio will be used by CCHMC researchers to narrow down the occurrence of possible full-forms with respect to the MeSH data.

2. Acronym Disambiguation Techniques

Researchers have developed two main approaches to address the problem of acronym disambiguation in the medical domain:

1. detecting and mapping abbreviations to their full-forms solely based on the content of the article [5-10], or
2. detecting and mapping abbreviations to the full-forms based on the data obtained from abbreviation databases [11, 12].

In our system we use a detection technique based on an abbreviation database. Instead of creating our own mechanism, we chose Medilexicon as it is the largest online database of pharmaceutical and medical abbreviations. None of the previously proposed methods meet the exact requirements specified by the CCHMC researchers; some of the approaches retrieve acronym/full-form pairs but do not use the Medilexicon database. None of the methods retrieves articles in which a selected acronym/full-form appears and filters them for MeSH information. Hence building a new system accessing Medilexicon and Pubmed was necessary.

Our system is customized to incorporate features of Pubmed and Medilexicon with the following restrictions:

1. Each letter of the abbreviation should be consecutive in the full-form.
2. We only need those articles that have MeSH headers but Pubmed retrieves all articles.
3. After having studied the output of both Pubmed and Medilexicon, what we need is a selected subset of the output of both (we do not need all the full-forms retrieved by Medilexicon and we do not need all the articles retrieved by Pubmed).
4. As of now these two websites show the details of only one article at a time and we need all the article details to be shown together.

The next section describes and illustrates our framework.

3. System Description and Illustration

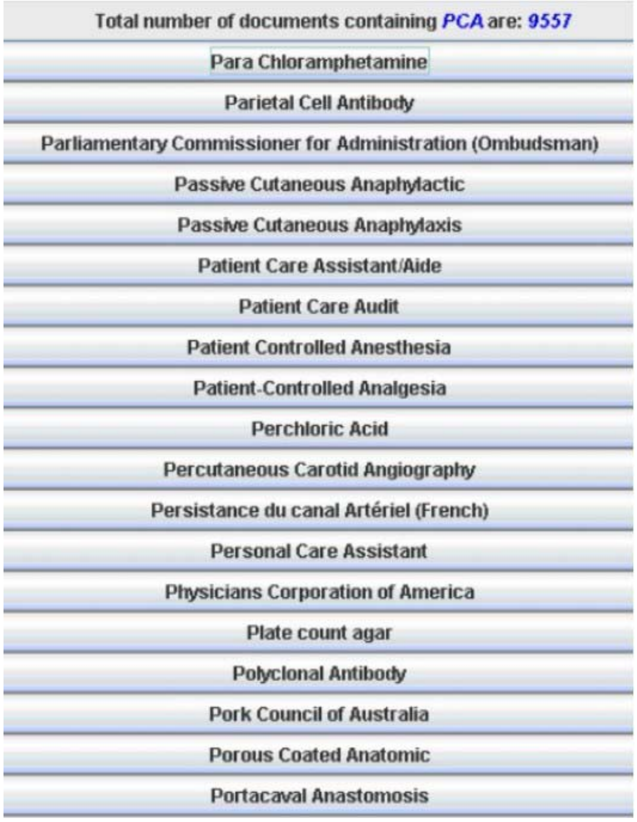
The system has three main functions:

1. Getting the abbreviation from the user and retrieving all of its full-forms from Medilexicon.
2. Letting the user select one full-form and retrieving all the articles containing that full-form from Pubmed.
3. Retrieving title, abstract, authors, journal, title, publication date, and MeSH headers from each article for the selected full-form, storing it in the database, and displaying the same to the user.

An excerpt from the output of Step 1 is shown in Figure 1 for the example acronym 'PCA.' The scroll bar is not shown, but a user can scroll through the full-forms and click on one to see the details associated with that term; our system accesses Pubmed to retrieve all articles containing the selected full-form and acronym.

MeSH headers for the selected full-form are shown in Figure 2. When a user clicks on a button the output shown in Figure 3 is displayed. Figure 3 gives the details extracted from Pubmed for the MeSH header 'Mice' and the acronym/full-form pair 'PCA' and 'Passive Cutaneous Anaphylactic,' respectively.

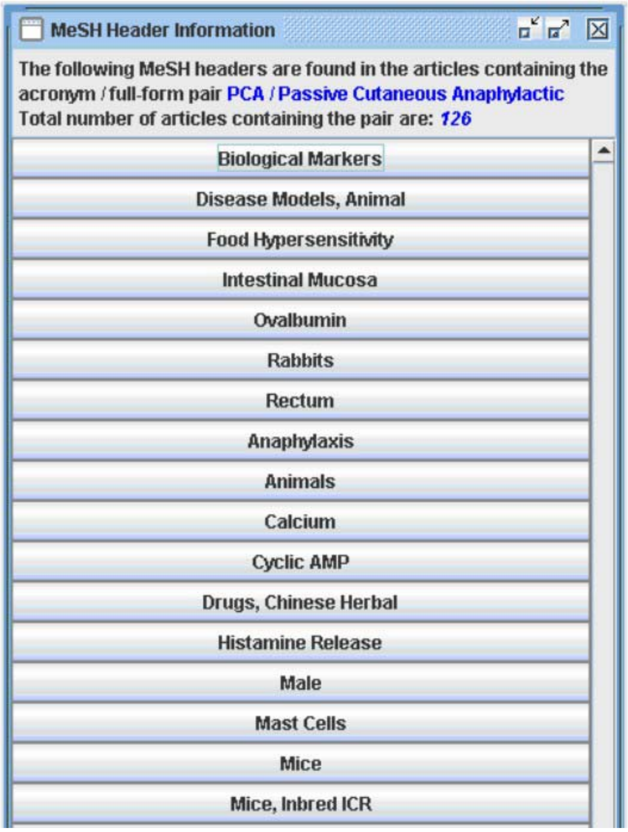
An overview of the entire system framework is given in Figure 4. The system performs the functions described in the following steps which are also shown in the figure:



Total number of documents containing *PCA* are: *9557*

Para Chloramphetamine
Parietal Cell Antibody
Parliamentary Commissioner for Administration (Ombudsman)
Passive Cutaneous Anaphylactic
Passive Cutaneous Anaphylaxis
Patient Care Assistant/Aide
Patient Care Audit
Patient Controlled Anesthesia
Patient-Controlled Analgesia
Perchloric Acid
Percutaneous Carotid Angiography
Persistance du canal Artériel (French)
Personal Care Assistant
Physicians Corporation of America
Plate count agar
Polyclonal Antibody
Pork Council of Australia
Porous Coated Anatomic
Portacaval Anastomosis

Figure 1. Full-forms of 'PCA' (Excerpt)

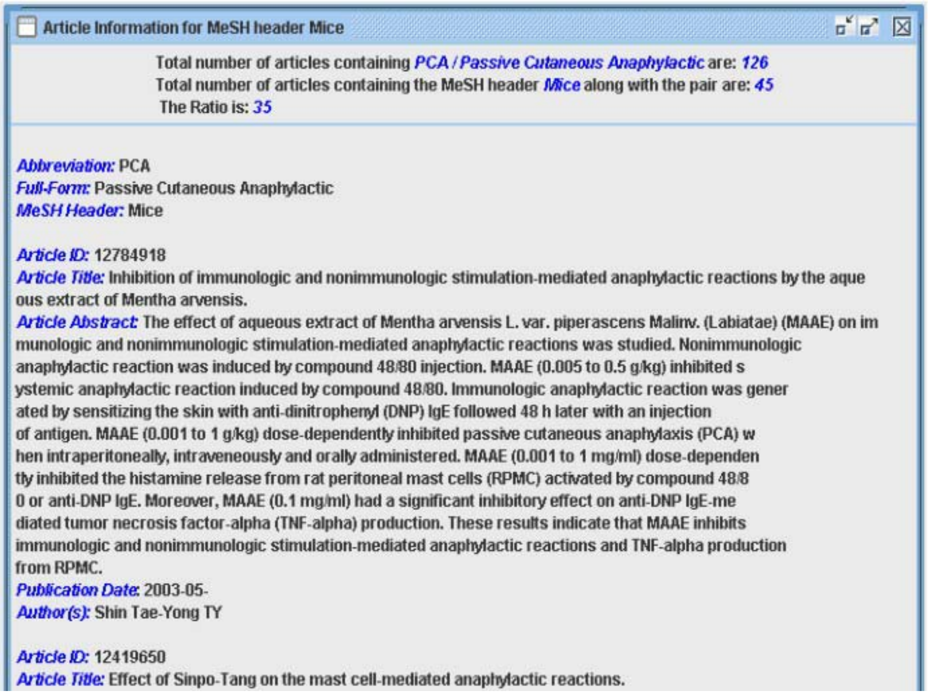


MeSH Header Information

The following MeSH headers are found in the articles containing the
acronym / full-form pair PCA / Passive Cutaneous Anaphylactic
Total number of articles containing the pair are: *126*

Biological Markers
Disease Models, Animal
Food Hypersensitivity
Intestinal Mucosa
Ovalbumin
Rabbits
Rectum
Anaphylaxis
Animals
Calcium
Cyclic AMP
Drugs, Chinese Herbal
Histamine Release
Male
Mast Cells
Mice
Mice, Inbred ICR

Figure 2. MeSH Headers (Excerpt)



Article Information for MeSH header Mice

Total number of articles containing *PCA / Passive Cutaneous Anaphylactic* are: *126*
Total number of articles containing the MeSH header *Mice* along with the pair are: *45*
The Ratio is: *35*

Abbreviation: PCA
Full-Form: Passive Cutaneous Anaphylactic
MeSH Header: Mice

Article ID: 12784918
Article Title: Inhibition of immunologic and nonimmunologic stimulation-mediated anaphylactic reactions by the aque
ous extract of Mentha arvensis.
Article Abstract: The effect of aqueous extract of Mentha arvensis L. var. piperascens Malinv. (Labiatae) (MAAE) on im
munologic and nonimmunologic stimulation-mediated anaphylactic reactions was studied. Nonimmunologic
anaphylactic reaction was induced by compound 48/80 injection. MAAE (0.005 to 0.5 g/kg) inhibited s
ystemic anaphylactic reaction induced by compound 48/80. Immunologic anaphylactic reaction was gener
ated by sensitizing the skin with anti-dinitrophenyl (DNP) IgE followed 48 h later with an injection
of antigen. MAAE (0.001 to 1 g/kg) dose-dependently inhibited passive cutaneous anaphylaxis (PCA) w
hen intraperitoneally, intraveneously and orally administered. MAAE (0.001 to 1 mg/ml) dose-dependen
tly inhibited the histamine release from rat peritoneal mast cells (RPMC) activated by compound 48/8
0 or anti-DNP IgE. Moreover, MAAE (0.1 mg/ml) had a significant inhibitory effect on anti-DNP IgE-me
diated tumor necrosis factor-alpha (TNF-alpha) production. These results indicate that MAAE inhibits
immunologic and nonimmunologic stimulation-mediated anaphylactic reactions and TNF-alpha production
from RPMC.
Publication Date: 2003-05-
Author(s): Shin Tae-Yong TY

Article ID: 12419650
Article Title: Effect of Sinpo-Tang on the mast cell-mediated anaphylactic reactions.

Figure 3. Article Information for a MeSH Header (Excerpt)

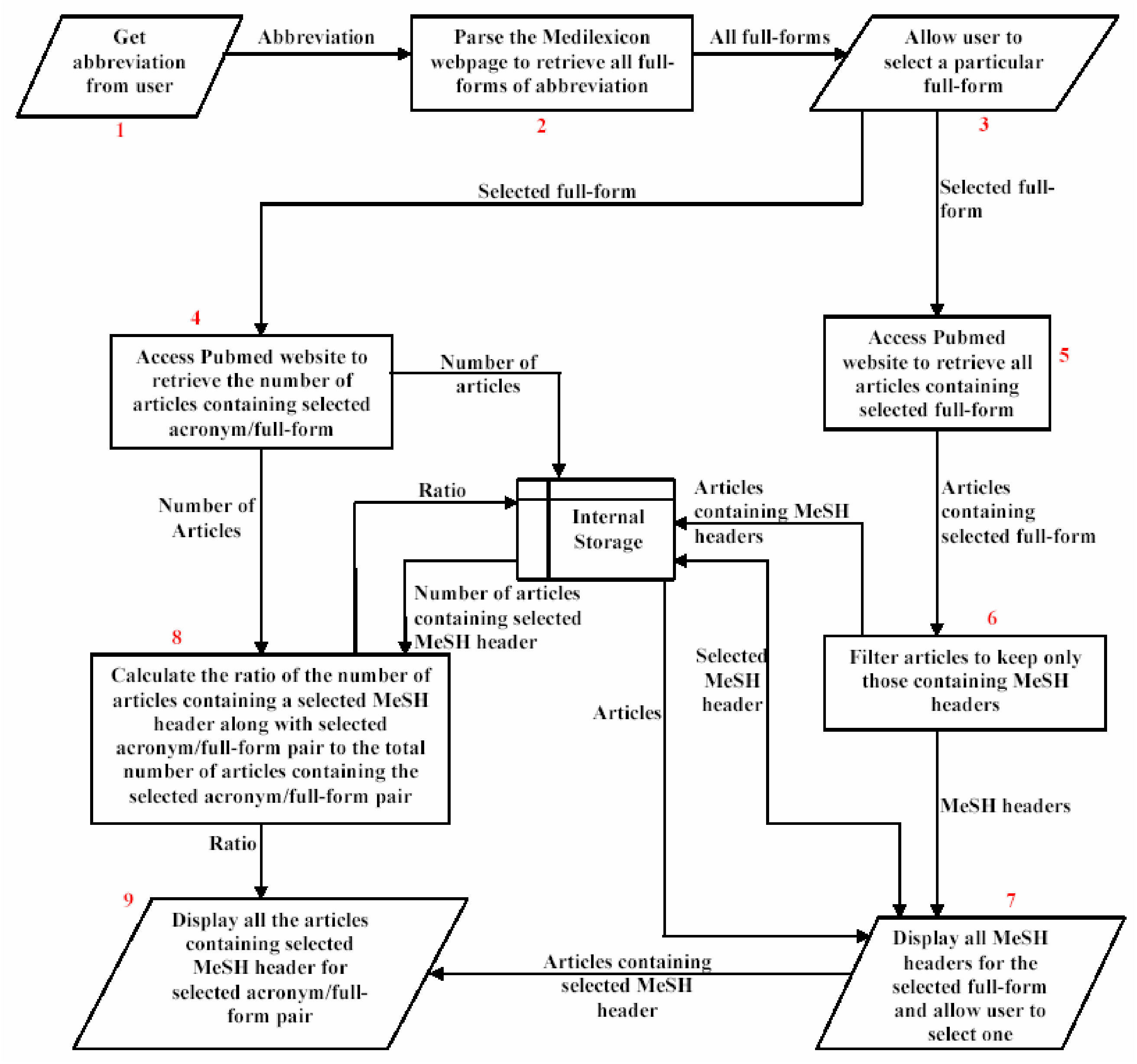

Get abbreviation from user
1
Abbreviation
Parse the Medilexicon webpage to retrieve all full-forms of abbreviation
2
All full-forms
Allow user to select a particular full-form
3
Selected full-form
Selected full-form
4
Access Pubmed website to retrieve the number of articles containing selected acronym/full-form
Number of articles
Access Pubmed website to retrieve all articles containing selected full-form
5
Number of Articles
Ratio
Internal Storage
Articles containing MeSH headers
Articles containing selected full-form
Number of articles containing selected MeSH header
6
Filter articles to keep only those containing MeSH headers
8
Calculate the ratio of the number of articles containing a selected MeSH header along with selected acronym/full-form pair to the total number of articles containing the selected acronym/full-form pair
Selected MeSH header
Articles
MeSH headers
Ratio
9
Display all the articles containing selected MeSH header for selected acronym/full-form pair
Articles containing selected MeSH header
7
Display all MeSH headers for the selected full-form and allow user to select one

Figure 4. A Framework for Medical Acronym Disambiguation

1. Display a GUI (Graphical User Interface) to interact with the user. The user will be a researcher working on medical text analysis who provides the abbreviation.
2. Create a connection to the Medilexicon webpage with the user given abbreviation as input and parse it to retrieve all full-forms of the abbreviation.
3. Display all the full-forms of the abbreviation to the user and provide the user with a facility to select a particular full-form.
4. Access the Pubmed website to retrieve the number of articles containing the abbreviation/full-form pair and store this count in the internal database.
5. Create a connection to the Pubmed website and retrieving all articles containing the selected full-form.
6. Parse the XML version of the article to check for MeSH header information, if the article has MeSH headers, store the article title, article abstract, publication date, author names and the MeSH headers in the internal database. If the article does not contain MeSH headers, it is discarded.

7. Display all MeSH headers for the selected full-form to the user and allow him/her to select one.
8. Calculate the ratio of the number of articles containing a selected MeSH header along with selected acronym/full-form pair to the total number of articles containing the selected acronym/full-form pair.
9. Display all the articles containing the MeSH header selected by the user in Step 7 and also display the ratio of occurrence of that MeSH header with respect to the acronym/full-form pair, as calculated in Step 8.

The next section describes the structure and contents of the internal database.

4. Database Design and Population

Our relational database has five tables (shown in Figure 5); primary keys are in boldface. Most of the information stored in these tables is retrieved by parsing XML pages. For example, the webpage containing the Pubmed article details is parsed to retrieve the article title, article abstract, names of the authors and MeSH header information. The publication date in the article may not always contain a specific date and may be missing the day or month and hence needs to be formatted before storing it into the database. For example, the date should consist of the day, month and the year and should appear as 09/12/2004, but sometimes it may be missing one of the three. In that case we insert '-' in place of the missing value.

Similarly the author names appear in the XML page in three parts, the first name, middle name, and last name. Sometimes the <FirstName> tag in the XML page is replaced by the <ForeName> tag, so we need to check for both while parsing XML page. Once these three parts are retrieved they are combined to form the full name of the author and this is stored in the database.

As we have seen from the publication date and author tags in the XML file these tags may vary or be missing, so we have a database table to store the XML tags for the information we need from the webpage and these tags are scanned for in the XML code when parsing it. A table *Tagname* (not shown here) stores tag mappings. Whenever a new tag for existing data is encountered in the XML code, e.g., <ForeName> instead of <FirstName>, we add the new tag to the table.

5. Acronym Full-form Disambiguation

The system described here is part of a larger research effort to investigate whether MeSH header information can be used effectively to disambiguate acronym/full-form pairs in medical text such as clinical annotations. In this section we use an example to demonstrate how the results generated by our system will be useful to the researchers at CCHMC.

Acronym table

abb_id	*abbreviation*	*total_docs*
1	SMA	4786
2	PCA	3592

Meshheader table

mesh_id	*desc_name*	*qual_name*
1	Algorithms	Method
2	Mice	0
3	Animals	0

Definition table

abb_id	*def_id*	*definition*	*docs_per_def*
1	1	Smooth Muscle Antibody	92
1	2	Spinal Muscular Atrophy	59
1	3	Standard Method Agar	111
2	4	Percutaneous Carotid Angiography	215

Def_art_mesh table

mesh_id	***def_id***	***article_id***
1	1	346279
2	1	563372
3	2	934766

Article table

article_id	*article_title*	*abstract*	*pub_date*	*author*	*journal title*
346279	Heart Vessels	Hyperplasia suppressor gene …	12-14-1993	Jiang GJ, Han M	Clinical and Experimental Allergy
563372	Prognostic value of auto-antibodies in the …	To investigate the prevalence of a group of different auto- …	04-21-1997	Al-Shukaili, Al-Jabri AA, Al-Mou	International Archives of Allergy and Immunology

Figure 5: Database Implementation

We illustrate how our application assists in determining whether MeSH header information can be used to disambiguate acronym/full-form pairs. First we retrieve information for the acronym/full-form pair, e.g., 'PCA,' and 'Passive Cutaneous Anaphylactic.' Output from the system is a list of MeSH headers and details of the articles they appear in along with the acronym/full-form pair. Our system also gives (1) a count of the number of articles containing every MeSH header in the articles containing the pair, and (2) the total number of documents containing the acronym/full-form pair. All of this information is stored in the internal database so that it can be obtained at a later time. We then compare the counts of each MeSH header occurring along with the acronym/full-form pair to the total number of documents that contain the acronym/full-form pair. This indicates how many times each MeSH header appears along with the pair.

An example of this comparison is shown in Table 1, where we give the MeSH header data, the total number of articles that contain the acronym/full-form pair, and the document specificity with regard to the MeSH header data. The document specificity refers to the count of the number of articles containing a specific MeSH header along with a particular acronym/full-form pair. The last column in Table 1 displays the percentage of articles that contain a specific MeSH header with respect to the total number of articles containing the acronym/full-form pair.

Table 1. 'PCA/Passive Cutaneous Anaphylactic'Articles

MeSH header descriptor data	**articles with acronym, full-form**	**articles containing given MeSH header**	**prcentage of MeSH header articles to total articles**
Mice	126	45	35.7%
Mice, Inbred Strains	126	31	24.6%
Mice, Mutant Strains	126	2	1.58%
Mice, Transgenic	126	1	0.79%
Mice, Inbred A	126	1	0.79%
Mice, Inbred BALB C	126	14	11.11%
Mice, Inbred C3H	126	3	2.38%
Mice, Inbred C57BL	126	3	2.38%
Mice, Inbred ICR	126	9	7.14%
Mice, Knockout	126	1	0.79%
Mice, Nude	126	1	0.79%

Table 1 displays results of a few MeSH headers of the 'Mice' hierarchy which appear in articles containing the pair 'PCA/Passive Cutaneous Anaphylactic.' If the number of articles containing a specific MeSH header along with the acronym/full-form pair equals the total number of articles containing the acronym/full-form pair, in this case 126, then we can claim that the MeSH header identifies the acronym/full-form pair with 100% accuracy. The MeSH header 'Mice' is contained in 45 documents whereas the acronym/full-form pair appeared together in 126 articles. We conclude that the MeSH header 'Mice' does not identify the acronym, full-form pair ('PCA,' 'Passive Cutaneous Anaphylactic') with 100% accuracy but by 35% accuracy with respect to the total number of articles containing the pair.

Our work provides a foundation for the researchers at CCHMC to conduct analysis to validate or invalidate the premise that the MeSH headers can disambiguate acronym/full-form pairs. From the small datasets we have worked with, it is inconclusive whether the MeSH header information is useful for helping reduce the ambiguity created by the use of acronyms. Further investigation is needed.

6. Considering All Full-forms

In this section we discuss some additional tests that we carried out for the researchers at CCHMC. For this phase of testing we removed the restriction that only a single full-form is selected. We retrieve and store in our internal database all the full-forms and the number of articles containing the acronym/full-form pair.

Once the user selects an abbreviation we retrieve all the MeSH headers found in the articles containing the selected acronym/full-form pair. When the user selects a MeSH header we retrieve all the articles containing the MeSH header along with the acronym and every full-form of the acronym. For example, we retrieve all the full-forms for the acronym, e.g., 'MIN' in Table 2. Once the user selects a full-form, 'Melanocytic Intraepidermal Neoplasia,' all the MeSH headers found in the articles containing 'MIN/Melanocytic Intraepidermal Neoplasia' are displayed to the user. If the user selects a MeSH header 'Melanoma,' our tool calculates the support and the confidence for each full-form of the acronym.

These terms can be defined as follows: *support* is the measure of how often a collection of items occur together as a percentage of all the transactions [13]. For example, the number of times the MeSH header 'Melanoma' occurs in articles containing both 'MIN' and 'Melanocytic Intraepidermal Neoplasia.' *Confidence* of rule '*B* given *A*' is a measure of how much more likely it is that *B* occurs when *A* has occurred. When used with association rules, the term confidence is observational rather than predictive [13].

The calculation of support and confidence provide further information to the researchers. Support gives the percentage of the number of articles containing a specific MeSH header along with the acronym/full-form pair. Confidence is the percentage of the number of articles containing a specific MeSH header along with an acronym/full-form pair with respect to the total number of

Table 2. Support and Confidence for All Full-forms

Full-Form	Number of documents containing MIN and full-form	Number of documents containing MIN/full-form and MeSH header Male	Support (col3/col2) * 100	Confidence (col3/total(col3))* 100; Total (col3) = 20899
Mammary Intraepithelial Neoplasia	10	2	20.00%	0.009%
Medial Interlaminar Nucleus	36	4	11.11%	0.020%
Melanocytic Intraepidermal Neoplasia	3	1	33.33%	0.005%
Microsatellite Instability	90	29	32.22%	0.010%
Microsatellite Instability Pathway	11	3	27.28%	0.004%
Minimal Model	1108	603	54.42%	2.890%
Minisatellite Instability	1	0	0.00%	0.000%
Minocycline	71	30	42.25%	0.140%
Minoxidil	91	64	70.33%	0.310%
Multifocal Intraepithelial Neoplasia	2	0	0.00%	0.000%
6-Multiple Intestinal Neoplasia	0	0	0.00%	0.000%
Minimal or Minute	**17089**	**9711**	**93.75%**	**46.470%**
Minute (French)	106	58	54.72%	0.280%
Minute (MIN = Mobile Identification Number)	0	0	0.00%	0.000%
Multiple Intestinal Neoplasia	151	80	52.98%	0.380%
Inter-Meal Interval	5	3	60.00%	0.014%
Mean (SD) Duration	1226	827	67.46%	3.960%
Mineral	2570	690	26.85%	3.300%
Minim	3	0	0.00%	0.000%
Minimum, Minimal	168	84	50.00%	0.400%
Minor	5247	2802	53.42%	13.400%
Minute	9580	5908	61.67%	28.270%
Minutos (spanish)	0	0	0.00%	0.000%

articles containing the specific MeSH header and acronym and any of the full-forms of the acronym. Table 2 shows an example dataset for the acronym 'MIN' and all its full-forms for the MeSH header 'Male.' A support value of 93.75% is shown in the highlighted row, which indicates that 93.75% of all articles containing the acronym/full-form pair 'MIN/Minimal or Minute' also contain the MeSH header 'Male.' The confidence value of 46.47% means that if an article contains the acronym 'MIN' and the MeSH header 'Male,' then there is a 46.47% chance that the full-form of 'MIN' that occurs in the article is 'Minimal or Minute' as compared to all other full-forms of 'MIN.'

7. Conclusions and Future Work

We offer three main contributions of our system in this paper. One, our system retrieves facts about the occurrence of an acronym along with its full-form. For example, we have found that 'PCA' occurs in 9557 articles, 'Passive Cutaneous Anaphylactic' occurs in 403 articles, and they both occur together in 126 articles. Hence, 'PCA' occurs by itself in 9431 articles and

‘Passive Cutaneous Anaphylactic’ occurs by itself in 277 articles. Considering only those medical articles that contain both the acronym and its full-form, we parse them to check for MeSH header information. Continuing with the example, we can use our system to find out that out of the 126 articles containing both, all 126 articles contain MeSHheader information. This information will be used by the CCHMC researchers in further research efforts to determine whether the MeSH header information can be used to disambiguate acronyms. For example, if a specific MeSH header is found in at least 113 articles out of 126, we can say that the MeSH header identifies the pair (‘PCA,’ ‘Passive Cutaneous Anaphylactic’) by 90% or more of the total number of articles containing the pair.

A second contribution of this system is that all the information required to disambiguate acronyms using MeSH headers has been consolidated into our internal database. Researchers do not need to refer to different websites and collect information manually; our system retrieves required information and provides it for reference at a later time. A user can view the MeSH header information for a specific acronym/full-form pair for which data has already been retrieved and stored in the database.

Our third contribution is that this tool is designed to be extensible, by which we mean that it is designed in a modular way so that new functionality can be added later. It is also designed to easily accommodate changes in parsing and retrieval of information from XML files. Currently, the tool is useful in any biological domain. It would be useful in other domains provided that an ontology and an acronym dictionary for that domain are defined.

Our work provides a foundation for various promising research and development initiatives, such as a research effort to see if MeSH header information can be used to disambiguate acronym/full-form pairs. The application illustrated in Section 6 retrieves information for a small sample dataset to give an idea of the kind of use it facilitates for bioinformatics research. As of now we retrieve information for MeSH headers present in the medical articles that contain an acronym, full-form pair. As future work, parsing the complete hierarchy tree for every MeSH header and retrieving information for all the MeSH headers in the tree would create a larger database and provide researchers with more information to use in their investigations. A web-based version of the tool with this extension is expected to be developed and made available to the research community by the CCHMC Computational Medicine Center.

8. References

[1] “Medical Subject Headings,” Service of United States National Library of Medicine and National Institute of Health, September 1, 1999; date accessed: March 2, 2006, http://www.nlm.nih.gov/mesh.

[2] D. Riecks, “Controlled Vocabulary,” date accessed: April 6, 2006, http://www.controlledvocabulary.com.

[3] “Pubmed Home Page,” Service of United States National Library of Medicine and National Institute of Health, October 10, 1993; date accessed: March 2, 2006, http://www.pubmed.gov.

[4] “Medilexicon,” Medilexicon International Ltd., date accessed: March 2, 2006, http://www.medilexicon.com.

[5] Schwartz, A.S., and Hearst, M.A., “A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text,” *Proceedings of the 8th Pacific Symposium on Biocomputing,* pages 451–462, Hawaii, January 2003.

[6] Yoshida, M., Fukuda, K., and T. Takagi, “PNAD-CSS: a Workbench for Constructing a Protein Name Abbreviation Dictionary,” *Bioinformatics,* Volume 16, Issue 2, pages 169-175, February 2000.

[7] Yu, H., Hripcsak, G., and C. Friedman, “Mapping Abbreviations to Full Forms in Biomedical Articles,” *Journal of American Medical Informatics Association,* Volume 9, Issue 3, pages 262-272, May 1, 2002.

[8] Bowden, P.R., Evett. L., and P. Halstead, “Automatic Acronym Acquisition in a Knowledge Extraction Program,” *Proceedings of the First Workshop on Computational Terminology, CompuTerm ’98,* pages 43-49, Montreal, Ontario, August 15, 1998.

[9] Taghva, K. and J., Gilbreth, “Recognizing Acronyms and Their Definitions,” *International Journal on Document Analysis and Recognition,* pages 191-198, 1999.

[10] Liu, H., and C., Friedman, “Mining Terminological Knowledge in Large Biomedical Corpora,” *Proceedings of the 8th Pacific Symposium on Biocomputing,* pages 415-426, Hawaii, January, 2003.

[11] Hahn, U., Daumke, P., Schulz, S., and K.G. Marko, “Cross-Language Mining for Acronyms and Their Completions from the Web,” *Proceedings of the 8th International Conference on Discovery Science,* pages 113-123, Singapore, October 8-11, 2005.

[12] Chang, J.T., Schutze, H., and R. B., Altman, “Creating an Online Dictionary of Abbreviations from MEDLINE,” *Journal of American Medical Information Association,* Volume 9, Issue 6, pages 612–620, November 1, 2002.

[13] “Data Mining Glossary,” Two Crows Consulting, date accessed: January 24, 2007, http://www.twocrows.com/glossary.htm.