TASK-DRIVEN DYNAMIC TEXT SUMMARIZATION

by

Terri Elizabeth Workman

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

December 2011

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Terri Elizabeth Workman**

has been approved by the following supervisory committee members:

John F. Hurdle	, Chair	**Oct 27, 2011** Date Approved
Marcelo Fiszman	, Member	**Nov 30, 2011** Date Approved
Stephane M. Meystre	, Member	**Oct 31, 2011** Date Approved
Joyce A. Mitchell	, Member	**Oct 26, 2011** Date Approved
Joan M. Stoddart	, Member	**Oct 26, 2011** Date Approved

and by **Joyce A. Mitchell** , Chair of

the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

The objective of this work is to examine the efficacy of natural language processing (NLP) in summarizing bibliographic text for multiple purposes. Researchers have noted the accelerating growth of bibliographic databases. Information seekers using traditional information retrieval techniques when searching large bibliographic databases are often overwhelmed by excessive, irrelevant data.

Scientists have applied natural language processing technologies to improve retrieval. Text summarization, a natural language processing approach, simplifies bibliographic data while filtering it to address a user's need. Traditional text summarization can necessitate the use of multiple software applications to accommodate diverse processing refinements known as "points-of-view."

A new, statistical approach to text summarization can transform this process. Combo, a statistical algorithm comprised of three individual metrics, determines which elements within input data are relevant to a user's specified information need, thus enabling a single software application to summarize text for many points-of-view. In this dissertation, I describe this algorithm, and the research process used in developing and testing it. Four studies comprised the research process. The goal of the first study was to create a conventional schema accommodating a genetic disease etiology point-of-view, and an evaluative reference standard. This was accomplished through simulating the task of secondary genetic database curation. The second study addressed the development

and initial evaluation of the algorithm, comparing its performance to the conventional schema using the previously established reference standard, again within the task of secondary genetic database curation. The third and fourth studies evaluated the algorithm's performance in accommodating additional points-of-view in a simulated clinical decision support task. The third study explored prevention, while the fourth evaluated performance for prevention and drug treatment, comparing results to a conventional treatment schema's output.

Both summarization methods identified data that were salient to their tasks. The conventional genetic disease etiology and treatment schemas located salient information for database curation and decision support, respectively. The Combo algorithm located salient genetic disease etiology, treatment, and prevention data, for the associated tasks.

Dynamic text summarization could potentially serve additional purposes, such as consumer health information delivery, systematic review creation, and primary research. This technology may benefit many user groups.

for Mary Ann

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my committee members for their sound advice and encouragement. They served as very willing and capable advisors in my efforts of combining the fields of biomedical informatics and information studies in addressing shared natural language processing issues. I especially wish to thank Dr. John Hurdle, my committee chair, for listening to my ideas, editing my written thoughts, and offering continual direction.

I want to thank the National Library of Medicine, for funding my research (grant number T15LM007123) and giving me the chance to work with their extraordinary scientists. I would not have succeeded in my research goals without the help of Marcelo Fiszman, Thomas Rindflesch, Gabriela Rosenblatt, Anna Ripple, and others.

Most of all, I want to thank my family, particularly my daughter and parents, for their faith and patience. I also want to thank my cat, for being a kind, warm, constant friend throughout this experience.

CHAPTER 1

INTRODUCTION

Objective and Hypothesis

The objective of this work is to examine the efficacy of natural language processing (NLP) in summarizing bibliographic text for multiple purposes. The central research hypothesis is that an NLP text summarization process that transforms bibliographic text into a topically filtered, compact form can be used to extract and identify data crucial to multiple information needs. This is dependent on the subhypothesis that, once it is transformed into a basic compact form, bibliographic text collectively retains the thematic focus that was expressed in the initial search query used to retrieve it. Because it retains this thematic focus, various types of analysis can be used to extract elements from the output which are salient to a specific information task. This can be demonstrated through applying text summarization to simulated tasks, and evaluating the summarized results with task-oriented methodologies and reference standards.

Motivation

The central motivation to this work is the continuing growth of bibliographic databases, and the problematic issues it creates. Researchers have documented the phenomenon of accelerated growth in bibliographic databases [1], which has created challenges to users practicing traditional information retrieval (IR) search techniques. These techniques, when applied to large bibliographic databases such as MEDLINE, can return a large, unmanageable list of citations, providing data that often do not fulfill the searcher's information needs [2]. One potential reason that traditional IR fails to meet a user's needs is because the user brings a "point-of-view" to the search that the IR engine either does not know or cannot exploit. A point-of-view is an additional concept

emphasis, such as treatment or diagnosis, which can be applied while locating data. It parallels subheading refinement, an available option in some controlled vocabularies like MeSH [3].

Aims

There are three aims for this research, each examining the use of text summarization for a specific task or tasks. To facilitate this, I performed the work using an information processing model called Semantic MEDLINE [4].

Aim 1: Develop and evaluate the effectiveness of Semantic MEDLINE in summarizing MEDLINE data for a new point-of-view, genetic etiology of disease, for the task of secondary database curation [5].

Aim 2: Develop and test a new algorithm that automatically identifies predications salient to a seed topic and the point-of-view expressed in a search query, within the domain of secondary database curation using the results of Aim 1[6].

Aim 3: Using the algorithm from Aim 2, create a dynamic summarization application and evaluate its performance for two additional points-of-view [7, 8].

Although I used Semantic MEDLINE as a test bed for this work, the specific methods can likely be applied in any other environment in which (a) initial text is converted to *subject_predicate_object* triplets, and (b) there is a sufficient database of triplets to form the general data profiles used in the algorithm computations described in this chapter.

Background

Natural Language Processing and Text Summarization

Elizabeth Liddy defines Natural Language Processing (NLP) as "a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [9]. It is an umbrella term that includes individual functions like part-of-speech tagging and sentence parsing, as well as more complex applications like information retrieval, machine learning, and information extraction [10]. It emerged as a topic in computer science in the years following World War II, through the work of pioneer scientists such as Alan Turing [11] and Noam Chomsky [12]. NLP models can often be divided into the two theoretical approaches of formal rule systems, and probabilistic models [10]. Formal rule systems, such as context-free grammars, model language behavior through rules dictating how language units like words and phrases are logically grouped. Probabilistic models like n-grams accomplish this through determining such groupings through statistical probabilities.

Text summarization is a natural language processing subdomain, which emerged in the late 1950's [13]. Its goal is to abstract relevant content from single or multiple sources [14]. Text summarization generally uses an extractive or abstractive approach. Extractive summarization provides verbatim chunks, or "extracts" of the original document(s), by appraising the lexical or statistical value of text units, or matching patterns of phrases. Abstractive methodologies summarize by describing the original content through paraphrasing or synthesizing the original text [15]. It usually relies on

outside knowledge sources. Text summarization generally produces indicative, informative, or critical summaries. The intent of indicative summaries is to simply alert and guide users to the original source documents, which then can be directly reviewed. Informative summaries identify the most salient content in source documents and present it in a structured form to the user as a surrogate for the original text. Critical summaries present informative content, along with some sort of a critical appraisal of the original documents. Text summarization can be evaluated using either intrinsic or extrinsic methods. In intrinsic evaluation, the quality of output is determined by analyzing the summary itself. Evaluators can appraise the fluency of the summary, compare it to an ideal summary of the same original content, or determine if it expresses previously chosen "key ideas". In extrinsic evaluation, summarization is judged by its value in completing a separate task.

While text summarization has been extensively used in the mass media domain (as noted by Zhang [16] and reflected in Inderjeet Mani's work [17]), it has also been applied in the biomedical domain. Yang and his associates clustered gene information using free text, MeSH, and Gene Ontology features, and presented summarizes based on sentence rankings [18]. Using an application called PERSIVAL, McKeown et al. retrieved, ranked, and summarized documents according to a patient's profile information [19]. Yoo et al. used an ontology-enhanced approach to cluster similar documents, then summarized the content within document groups by building text semantic interaction networks using semantic relationships within the document clusters [20]. Cao and his colleagues used a machine learning approach to classify questions, and also utilized a clustering technique using query keywords for presenting output, for their AskHERMES

system [21]. Text summarization applications such as Semantic MEDLINE that utilize semantic predications have the advantage of presenting an abstracted, compact expression of the original information that can be filtered according to a user's specific information need. Semantic predications are succinct *subject_verb_object* declarations that simplify the meaning of the PubMed text from which they are drawn. Due to their structure, they are subject to computational analysis.

Semantic MEDLINE

Semantic MEDLINE [4], developed at the National Library of Medicine (NLM), is a multistage natural language processing model designed to extract meaningful information from biomedical bibliographic citations. It is a summarization application in the abstraction paradigm, and relies on the Unified Medical Language System (UMLS) [22] knowledge source. The user initiates use of the Semantic MEDLINE application by submitting a search query expressing his or her information need. Semantic MEDLINE then relies on the separate, sequential applications of SemRep, Summarization, and Visualization to (respectively) transform the citations' title and abstract text into a compact form, identify resulting data which are salient to a specific information need, and display the results in a graphic, visual format. The following text describes these processes in detail.

SemRep

SemRep, a rule-based symbolic NLP tool developed by Rindflesch [23], extracts meaning from PubMed citations in the form of *semantic predications*. Predications are a

distillation of information contained in a phrase or a sentence, and they are expressed as a triplet: *subject_predicate_object*. SemRep tokenizes the input citations' title and abstract text. It identifies and tags each word's part-of-speech using the MedPost tagger, [24] along with the UMLS SPECIALIST Lexicon to disambiguate vague terms. It then performs an underspecified parsing of the text, and maps nouns phrases using MetaMap technology [25]. SemRep uses indicator rules to map syntactic elements to predicates in the UMLS Semantic Network. Using logical constraints within the Semantic Network, SemRep builds the output semantic predications by identifying the rational relationship or predicate that binds the connected subject and object arguments. For example, SemRep transforms the following title text:

> "**Taurolidine** is effective in the **treatment** of central venous catheter-related bloodstream **infections** in cancer patients" [26]

into the following semantic predication:

> Taurolidine_TREATS_infection

SemRep identifies "taurolidine" and "infections" as the respective subject and object of the text, and maps them to the UMLS [22] Metathesaurus preferred concepts *Taurolidine* and *infection*. It also recognizes "treatment" as the relational concept binding the subject and object terms, mapping it to the predicate TREATS. SemRep also identifies the logical UMLS semantic group classifications associated with the arguments, which in this case are "Pharmacologic Substance" (associated with *Taurolidine*) and "Disease or Syndrome" (associated with *infection*).

Summarization

The Summarization phase, developed by Fiszman, [27] identifies SemRep output which is relevant to a specific, user-indicated topic. This process begins by prompting the user to select a specific UMLS metathesaurus concept from among those occurring in the SemRep output which most precisely expresses the topic associated with the user's information need. Once the user identifies the concept, Summarization processes the SemRep output with four sequential filters known as *Relevance, Connectivity, Novelty,* and *Saliency*:

Relevance: Gathers semantic predications containing the user-selected seed topic. For example, if the seed topic were *Endometrial carcinoma*, this filter would collect the semantic predication cetuximab-TREATS-Endometrial carcinoma.

Connectivity: Augments *Relevance* predications with those which share a nonseed argument. For example, in the above predication cetuximab-TREATS-Endometrial carcinoma, this filter would augment the *Relevance* predications with others containing *cetuximab*.

Novelty: Eliminates vague predications, such as pharmaceutical preparation-TREATS-patients, that present information that users already likely know, and are of limited use.

Saliency: Limits final output to predications that occur with adequate frequency. For example, if cetuximab-TREATS-Endometrial carcinoma occurred enough times, all occurrences would be included in the final output.

In order to process data, the functionality of these four filters is programmed into a software application.

In this dissertation's research, I have employed two approaches to Summarization, which I will refer to as conventional summarization and dynamic summarization. Conventional summarization relies on specified *subject_predicate_object* patterns to identify the optimal predicates and semantic type subject and object arguments allowed as final Summarization output, in order to capture data relevant to a given point-of-view. For example, note the following named groups of semantic types, and their placement as subject or object arguments to specific predicates, which express a genetic etiology of disease point-of-view:

Named Semantic Type Groups

Genetic phenomenon: Amino Acid Sequence; Enzyme; Genetic Function; Nucleic Acid, Nucleoside, or Nucleotide; Nucleotide Sequence; Amino Acid, Peptide, or Protein; Gene or Genome; and Molecular Sequence.

Anatomy: Anatomical Structure; Body Part, Organ, or Organ Component; Cell; Cell Component; Embryonic Structure; Fully Formed Anatomical Structure; Gene or Genome; and Tissue.

Disease Process: Acquired Abnormality; Anatomical Abnormality; Congenital Abnormality; Cell or Molecular Dysfunction; Disease or Syndrome; Injury or Poisoning; Mental or Behavioral Dysfunction; Neoplastic Process; Pathologic Function; Sign or Symptom; Biologic Function; Cell Function; Mental Process; Molecular Function; Natural Phenomenon or Process; Organism Function; Organ or Tissue Function; Physiologic Function; Behavior; Mental or Behavioral Dysfunction; and Finding.

Named Groups Serving as Arguments for Specified Predicates

1) {genetic phenomenon } AFFECTS {disease process}
2) {genetic phenomenon } AUGMENTS {disease process}
3) {genetic phenomenon } DISRUPTS {disease process OR anatomy}
4) {genetic phenomenon } ASSOCIATED_WITH {disease process}
5) {genetic phenomenon } PREDISPOSES {disease process}
6) {genetic phenomenon } CAUSES {disease process}
7) {genetic phenomenon } STIMULATES {genetic phenomenon }
8) {genetic phenomenon } INHIBITS {genetic phenomenon }
9) {disease process} COEXISTS_WITH {disease process}

Conventional summarization requires prior research to determine which predicates and semantic type arguments capture salient data for the given point-of-view. A separate software application is required in order to summarize data for each desired point-of-view.

Dynamic summarization utilizes a statistical algorithm that analyses the properties of each SemRep output dataset it receives as input. Various metrics calculate term frequencies in order to determine which semantic predications are salient to the user's selected UMLS Metathesaurus preferred concept. This enables Summarization to adapt to the characteristics of each dataset it processes, thus enabling summarization for multiple points-of-view using a single software application, without relying on restricted *subject_verb_object* patterns. The concept of dynamic summarization (within the Semantic MEDLINE model) and its mechanisms were created as part of this

dissertation's work. I describe in detail the Combo algorithm, the central point of Aim 2's work, later in this chapter.

Successful summarization can validate the central hypothesis that an NLP text summarization process that transforms bibliographic text into a topically filtered, compact form can be used to extract and identify data crucial to multiple information needs. Summarization output can be evaluated through simulating human tasks, and comparing results to gold standards of desired output. Successful dynamic summarization can validate the subhypothesis that once it is transformed into a basic compact form, bibliographic text collectively retains the thematic focus that was expressed in the initial search query used to retrieve it. To test this hypothesis, SemRep output originating from PubMed queries expressing multiple topics and points-of-view can be processed by the four sequential Summarization filters, with the Combo algorithm acting as the operational mechanism in the *Saliency* filter. This output can also be applied to simulated tasks, using reference standards to evaluate the results.

Visualization

Visualization [4] presents the summarized semantic predications in an interactive graph. The graph's central node is the seed topic. Arcs representing predicate relationships connect the seed topic node to other argument nodes. Users may click on an arc for information regarding the associated semantic predications. For example, users could click on a TREATS arc connecting the seed topic *Endometrial carcinoma* node to the *Laparotomy* node to find title and abstract citation text concerning the treatment of endometrial carcinoma and laparotomy. Users can also view the citation record in

PubMed, and possibly access the fulltext article. Users may also limit which relational arcs the graph displays. In Figure 1, the user has limited the displayed arcs to the TREATS predicate relations, and has clicked on the arc connecting *Hysterectomy* to the central concept node *Endometrial Carcinoma* in order to review citations addressing hysterectomy as a treatment option for endometrial carcinoma.

Work to Achieve Each Aim

Aim 1

Motivation

Secondary genetic database curators are challenged by an overabundance of data resulting from constantly evolving biotechnologies [28] and the growing volume of published findings [1]. Aim 1 was motivated in part by this curation dilemma as well as a desire to explore how Semantic MEDLINE, implementing a conventional summarization approach, addressed it. The work of Aim 1 provided a reference standard which served to evaluate the work of both Aims 1 and 2, and a conventional summarization software application that also served an evaluative purpose in Aim 2.

Methods

As earlier noted, in using a conventional summarization approach in Semantic MEDLINE, the user specifies both a seed topic and an explicit point-of-view. For example, a user could seek information addressing the diagnosis (a point-of-view) of coronary artery disease (a seed topic). Using this conventional methodology, there is a

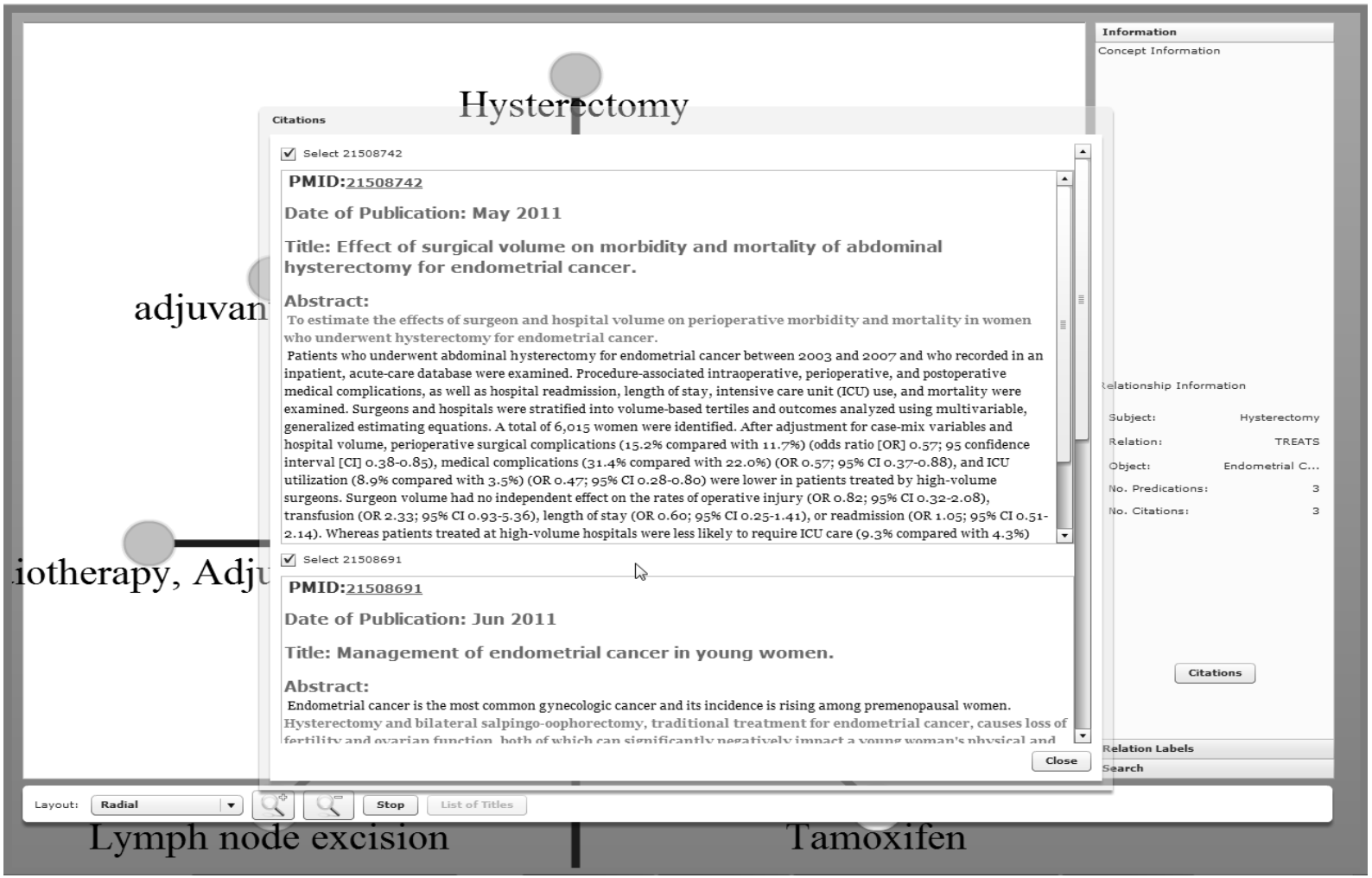

Figure 1. Visualization

limited number of points-of-view available to the user. Each software application facilitates summarization for a specific point-of-view, and must include a handcrafted set of specific, restrictive *subject-verb-object* patterns, but creating and evaluating such an application is nontrivial. Before completing the work of Aim 1, conventional summarization point-of-view options consisted of: treatment of disease [29]; substance interaction [30]; diagnosis [31]; and pharmacogenomics [32].

A software application developed in this Aim implemented the point-of-view of "the genetic etiology of disease," and built on the work of Rindflesch [33], Libbus [34], et al. Earlier they had identified the *predicates* expressing genetic disease etiology assertions in SemRep output. With Marcelo Fiszman's guidance, I assembled groups of semantic types which served as subject and object arguments. The software application I developed used these predicates and semantic type arguments, within the four-tiered summarization filtering framework, to harvest predications asserting a genetic etiology of disease point-of-view. I developed the software using Perl [35], an interpreted, high-level programming language. The other conventional summarization applications (e.g., treatment of disease, diagnosis), to which I had access were also developed using Perl. The treatment of disease application ably served as a framework and example for development of the genetic disease etiology software. The application also made use of the MySQL Semantic MEDLINE database for novelty processing.

Evaluation

To evaluate the new application's effectiveness we downloaded MEDLINE citations for SemRep processing using this query:

urinary bladder neoplasms[mh] OR "bladder cancer" OR "cancer of the bladder" Search output was limited to citations in English, with abstracts, that represented literature which was published from 1 January 2003 to 31 July 2008.

The citations were sequentially processed with SemRep and the summarization software. From the SemRep output, the summarization software identified semantic predications salient to the genetic etiology of bladder cancer, using *Carcinoma of bladder* as the seed topic for summarizing.

For evaluation, I assembled a reference standard of genes implicated in bladder cancer. I identified genes noted in Genetics Home Reference (GHR) [36] and Online Mendelian Inheritance in Man (OMIM) [37] records, based on source data from our study's timeframe (1 January 2003 to 31 July 2008). In order to find genes implicated in bladder cancer development as reported in OMIM, I retrieved records that were either phenotypically relevant to the disease, or provided a clinical synopsis, by executing the following search query:

"bladder cancer"[All Fields] OR "bladder cancers"[All Fields] OR "bladder cancer cases"[All Fields] OR "bladder cancer cell"[All Fields] OR "bladder cancer patients"[All Fields] OR "bladder carcinoma"[All Fields] OR "bladder carcinogenesis"[All Fields]

The query was executed twice. For the first execution, limits were adjusted in order to retrieve a broad range of genetic information addressing bladder cancer. The second execution focused exclusively on clinical synopses. Dr. Fiszman guided me in the search strategy. To locate relevant records in GHR, I searched using the keyword "bladder" to locate 11 relevant records. In order to build the reference standard, I manually reviewed

the OMIM and GHR records and listed genes with disease implications. I noted 10 significant genes in GHR records, and seven in OMIM records (with four of these genes present in both sources). The reference standard genes were compared to genes noted in the summarized output as appearing as subject arguments in semantic predications that featured the UMLS Metathesaurus concepts *Carcinoma of bladder*, *Bladder Neoplasm* or *Carcinoma, Transitional Cell* as object arguments. The gene subject arguments were also compared to their corresponding Entrez Gene record in measuring precision. If a gene argument did not appear in the reference standard, but its Entrez Gene record indicated it was implicated in bladder cancer development, it received a true positive status for precision.

The standard metrics of recall, precision, and F-score provided calculations to evaluate results. Recall consisted of the percentage of all reference standard items that were found in system output. Precision consisted of the percentage of system output gene arguments that were true positives. F-score was computed using the following function:

$$f(x) = 2(\text{recall} * \text{precision} / \text{recall} + \text{precision})$$

Results

PubMed produced 5606 citations. Using these as input, SemRep produced 38,498 semantic predications. Of these, the summarization software application identified 359 as salient. The summarization application achieved 0.46 recall, 0.88 precision, and an F-

measurement of .061 in comparing summarized gene association findings to the reference standard, reinforced with Entrez Gene data.

Chapter 2, which is also the text of an article [5] published by the Journal of the Medical Library Association, provides a detailed description of this work.

Aim 2

Motivation

Work for this aim was motivated by recognizing the need for a more adaptive summarization process, one unconstrained by the limited number of static points-of-view in Semantic MEDLINE. I developed and evaluated an algorithm for identifying salient semantic predications. It analyzes relevant attributes in SemRep output data with adapted statistical methods that have been successfully applied to other natural language processing tasks. I integrated the Kullback-Leibler Divergence (KLD) [38] and the RlogF [39] metrics to assess predicate and nonseed topic semantic type properties in SemRep output in order to identify the most significant semantic predications in a dataset. These metrics were combined with a scaling factor to form an algorithm called *Combo*. The Combo algorithm was evaluated for its effectiveness in identifying salient semantic predications, when acting as the computational mechanism in the *Saliency* filter in Semantic MEDLINE's summarization sequential four-tier architecture.

Methods

Algorithm Development

I investigated many metrics commonly used in natural language processing in developing the algorithm. This included basic relative frequency assessment [40], multiple inverted term frequency metrics [41], and a G^2 function used by Mani and Bloedorn [42]. After much research, I concluded that the three combined metrics noted above provided the most accurate statistical assessment of semantic predications. The following paragraphs give detailed descriptions of these metrics.

Previous research has noted a primary role of predicates in SemRep data in expressing a specific point-of-view [32, 33]. The Kullback-Leibler Divergence (KLD) measurement expresses the divergence between a true distribution (P) and an assumed distribution (Q). It has been successfully applied to prior NLP studies analyzing predicates [43]. When applied to predicate assessment, KLD accounts for superfluous predicates in SemRep output, rewarding the truly informative predicates by assigning to them higher scores. I hypothesized that a properly formed PubMed query contains a seed topic and point-of-view focus. The set of predicates from such a query, P, is compared to a set of predicates from a naively formed query, Q. The difference between the queries is that a properly formed query will include a MeSH subheading and possibly other details to adequate specify a point-of-view. For example, a naïve query for breast cancer treatment would be "Breast Cancer," while a properly formed query would be "Breast Neoplasms/therapy[majr]". The KLD measurement determines the collective difference between the two distributions, P and Q:

$$D(P||Q) = \Sigma\ P(x)\log_2(P(x)/Q(x))$$

The individual KLD calculations (before summing) for shared predicates serve as a means to determine which individual predicates are significant in expressing the intended point-of-view in SemRep data. By applying the KLD measurement exclusively to compare the relative frequency of individual predicates emerging from the properly formed query (distribution P) to their counterparts emerging from the naive query (distribution Q), one may calculate a score for each predicate representative of the proper query that indicates its importance in expressing the intended point-of-view. These scores may also be ranked to indicate which predicates are more influential in expressing the intended point-of-view initially expressed by the proper query.

Due to UMLS constraints, SemRep is limited in what concepts (and their matching semantic types) can be bound to a given predicate, in forming logical semantic predications. Therefore, semantic types tend to cluster with predicates in SemRep output. Such prominent associations express a predominant concept in the data, limited within the realm of each individual predicate. The RlogF measurement was developed by Riloff to assess the relevance of extracted patterns consisting of a syntactic constituent (i.e., a noun or verb phrase) and their arguments (i.e., a direct or indirect object), in information extraction processes. The RlogF measurement weighs an extracted pattern's conditional probability with the log of its frequency. I used RlogF to assess the value of a semantic type's "binding" to a given predicate. The RlogF measurement is expressed thus:

$$\text{RlogF}(\text{pattern}_i) = \log_2(\text{semantic type frequency}) * P(\text{relevant} \mid \text{pattern}_i)$$

$Pattern_i$ refers to a given predicate/semantic type binding, and the conditional probability ($P(relevant \mid pattern_i)$) is the quotient of the semantic type's raw frequency as bound to the predicate, divided by the raw frequency of all semantic types as bound to the same predicate. Dr. Hurdle suggested the use of this metric.

In semantic predication analysis, the magnitude of raw RlogF scores can exceed raw KLD scores, yet they express a different proportional relationship in SemRep output. KLD scores express a proportional relationship among predicates across the entire dataset, while RlogF scores express a binding between a single predicate and its associated semantic types. I developed a mechanism named PredScal to dynamically scale RlogF values according to the spatial proportions of predicates in a given dataset:

$$1 / \log_2(c)$$

where c represents the count of unique predicates in a dataset. In rare instances where there is only one unique predicate, the PredScal defaults to a value of 1. The three metrics are combined into a product called "Combo" to evaluate SemRep data distributions:

$$(RlogF * PredScal) * KLD$$

Data

MEDLINE citations returned by PubMed for the following search query were downloaded:

Urinary Bladder Neoplasms/genetics[majr] AND Urinary Bladder Neoplasms/etiology[majr] Language: English. ("2003/01/01"[Publication Date] : "2008/07/31"[Publication Date])

Analysis with Combo Algorithm

I analyzed the resulting semantic predications, using the Combo algorithm within the Semantic MEDLINE four-tier architectural model, with Combo implementation after manual *Relevance*, *Connectivity*, and *Novelty* filtering. To evaluate predications initially identified in the *Relevance* tier, I separated all semantic predications which included the UMLS Metathesaurus concept "Carcinoma of bladder" as either a subject or object argument. Semantic predications which were not considered novel were removed from this group, thus simulating the *Novelty* tier functionality in the model. Combo values were calculated as explained above for the semantic predications in this group (examining predications containing "Carcinoma of bladder" for KLD assessment). For evaluative purposes, the novel *Relevance* predications with the top four scores were considered salient.

The *Connectivity* tier augments novel *Relevancy* predications with others which share a nonseed topic semantic type as an argument. In order to examine salient *Connectivity* semantic predication identification, we performed a similar analysis on the SemRep output which did not include the seed topic "Carcinoma of bladder," but did share the nonseed concept semantic type "gene or genome" of the two highest scoring novel *Relevancy* predications. Such semantic predications were identified and then filtered so that only novel predications were included. Combo scores were calculated. It should be

noted that for calculating the KLD portion of the algorithm for connectivity filtering, predications containing the nonseed concept semantic type “gene or genome” were compared against their counterparts in the Semantic MEDLINE database. For evaluative purposes, novel *Connectivity* predications with the top score were considered salient.

Analysis with Traditional Summarization

For evaluative purposes, the same SemRep output was also processed with the conventional summarization software application created in Aim 1.

Evaluation

To evaluate the algorithm’s performance, I compared the results of the analysis using the Combo technique described above, to the conventional genetic disease etiology summarization software’s performance, utilizing the reference standard created for Aim 1. I measured results in terms of recall, precision, and F-score, using the same definitions of these metrics as noted in the work of Aim 1.

Results

The search query yielded 667 citations. SemRep processing produced 5,421 semantic predications. Summarizing the SemRep output using Combo resulted in 201 salient semantic predications, whereas the conventional software application identified 112 such predications. Combo identified 74 genetic entities implicated in bladder cancer development; the conventional application identified 10 implicated genetic entities.

To compare the effectiveness of the two summarization approaches, recall and precision values were calculated by comparing each set of genetic entities to the

reference standard developed for Aim 1. Summarization utilizing the Combo algorithm resulted in a 0.69 recall rate, whereas the conventional software application achieved a 0.23 recall rate. The Combo analysis achieved 0.81 precision; the conventional application achieved 1.0 precision. These calculations produced an F-score of 0.75 for the Combo method, and an F-score of 0.37 for the conventional application's output.

Chapter 3, an article [6] published by BMC Medical Informatics and Decision Making, details the work of Aim 2.

Aim 3

Motivation

A dynamic summarization application utilizing the Combo algorithm could potentially serve multiple needs. Building the application and analyzing its performance for a previously unaddressed information need, as well as an additional information need served through conventional summarization, would provide further insight to its intrinsic generalizability.

Methods

Application Development

The dynamic summarization software utilizes the same general four-tier architecture previously used in summarization. Figure 2 illustrates data flow within this architecture. The new Combo algorithm was incorporated into the final *Saliency* filter. This resulted in a new dynamic summarization application that transformed Semantic MEDLINE into

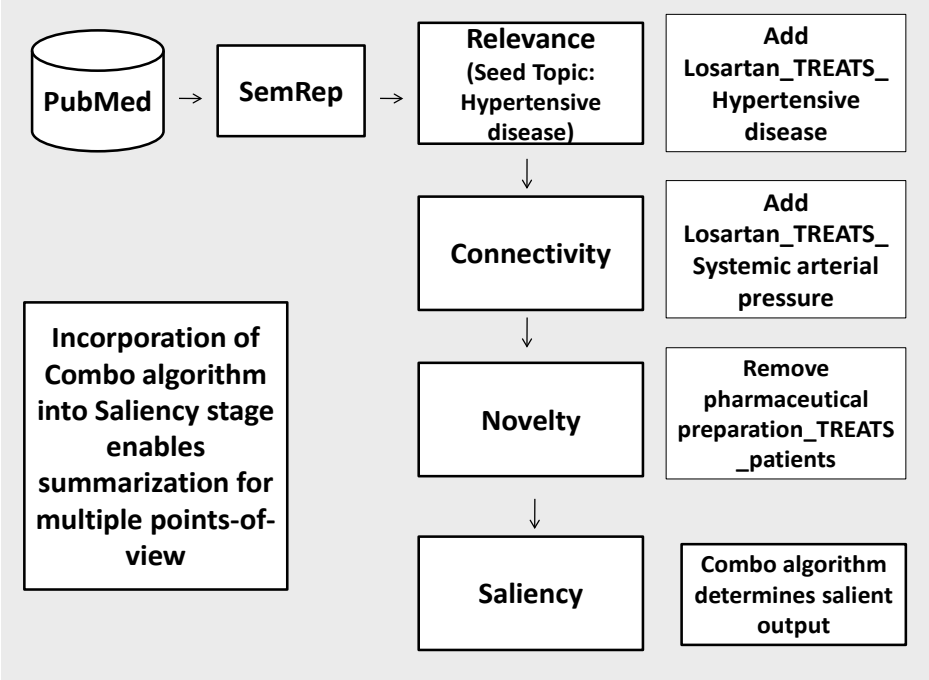

Figure 2. Dynamic Software Architecture

a generalizable, multipurpose application. Prior to this, conventional summarization allowed Semantic MEDLINE to process SemRep data for only five static points-of-view. The new dynamic software enabled Semantic MEDLINE to summarize for potentially many information needs.

The four Novel Relevance and four Novel Connectivity semantic predications with the highest Combo scores constituted Saliency output.

The dynamic software was developed using the Perl programming language. It utilized the MySQL Semantic MEDLINE database for novelty processing and KLD computations.

Application and Evaluation

We evaluated the dynamic software's performance in two additional points-of-view domains, specifically treatment and prevention. This facilitated evaluation for another point-of-view which has an established outcome using conventional summarization [29], and a point-of-view which is not presently served by conventional summarization.
A pilot project focused exclusively on prevention examined a hypothesis regarding expression of preventive interventions in semantic predications. This pilot project provided foundational work for a larger study exploring dynamic summarization for both points-of-view. We selected suitable disease topics to serve as subjects in each study. The pilot project examined the efficiency of dynamic summarization in locating preventive interventions for acute pancreatitis, malaria, and coronary artery disease. The second study examined the efficacy of dynamic summarization in locating preventive and drug treatment interventions for arterial hypertension, congestive heart failure, diabetes mellitus type 2, and pneumococcal pneumonia. We evaluated the results in both studies with reference standards consisting of DynaMed [44] recommendations. We also implemented a baseline methodology to evaluate the results of the second study.

Study One: Prevention Decision Support Pilot Project

Motivation. Online biomedical databases such as PubMed can be useful for patient care, yet users encounter obstacles in their effective use. [45] Suggested improvements include transforming text to provide clarified, explicit information. [46] Dynamic summarization may assist clinicians in identifying recommended disease prevention interventions. Could an NLP application such as Semantic MEDLINE, with the new

dynamic summarizing software, identify drug treatment options that are also cited in a decision support tool such as DynaMed?

Methods. I chose three disease topics: acute pancreatitis, malaria, and coronary artery disease. I formed and executed PubMed search queries using the following MeSH subject headings:

- Pancreatitis
- Coronary artery disease
- Malaria

combined with the *prevention and control* subheading. I processed the PubMed output with SemRep, and the dynamic summarizing software, using relevant disease seed topics for summarization. We built a reference standard for each disease by listing its recommended preventive interventions as found in DynaMed.

We assessed summarized output using primary and secondary evaluations. In the primary evaluation, we examined output in the form "Intervention X_PREVENTS_Disease Y", where the object argument was an expression of the topic disease. If the subject argument was one of the reference standard interventions, that intervention received a true positive status. If the subject argument was potentially a general expression of a reference standard intervention, we examined the full abstract and title text associated with the predication. If the precise reference standard intervention was named in the title or abstract, and the content identified it as a viable preventive intervention, it received true positive status. Recall, precision, and F-scores were calculated for each disease topic. Recall consisted of the percentage of reference standard interventions found in summarized output. I computed precision by grouping

subject arguments by name, and calculating what percentage was associated with a reference standard intervention. I used the same function used in the work of aims 1 and 2 to determine F-score.

In the secondary analysis, we examined all the other semantic predications which were not in the form "Intervention X_PREVENTS_Disease Y". We had hypothesized that preventive interventions could also be expressed in predications of other forms. The purpose of the secondary analysis was to explore this hypothesis. If additional reference standard interventions were found, recall and F-score would be recalculated by adding these new findings.

Results. The three PubMed search sessions retrieved a total of 3276 citations. SemRep produced a total of 19154 semantic predications. Summarization yielded 1964 semantic predications. The primary analysis resulted in 0.70 average recall, 0.45 average precision, and an overall F-score of 0.54. Additional reference standard interventions appeared in the secondary analysis, in specific forms such as "Intervention _USES_Intervention" (e.g., Prophylactic treatment_USES_Amodiaquine), "Intervention_TREATS_Person" (e.g. Malaria Vaccines_TREATS_Child), and "Intervention_TREATS_disease" (e.g., Secondary prevention_TREATS_Coronary arteriosclerosis, with "Secondary prevention" referencing smoking cessation), thus validating the hypothesis. We recalculated recall, resulting in a modified average recall of 0.79 and a recalculated overall F-score of 0.57.

The details of this study are included in a manuscript accepted by the Journal of the Medical Library Association, which also serves as Chapter 4 of this dissertation.

Study Two: Preventive and Drug Treatment Intervention Decision Support

Motivation. As earlier stated, online databases such as PubMed can potentially provide decision support information for patient care, yet obstacles render their use impractical. The dynamic summarization software may assist clinicians in identifying preventive and drug treatment interventions. Could an NLP application such as Semantic MEDLINE, with the new dynamic software, identify interventions that are also cited in a decision support tool such as DynaMed?

Methods. We chose four disease topics to serve as subjects for both prevention and drug treatment. These disease topics were arterial hypertension, congestive heart failure, diabetes mellitus type 2, and pneumococcal pneumonia. Dr. Meystre suggested these disease topics. I selected the following MeSH headings for these diseases:

- Hypertension
- Diabetes mellitus, type 2
- Heart failure
- Pneumonia, pneumococcal

I combined these MeSH headings with the subheading *drug therapy* to retrieve citations focused on drug treatment. I also combined the same MeSH headings with the subheading *prevention and control* to retrieve citations focused on prevention. The resulting citations were processed with SemRep, and then the dynamic summarizing software, using relevant seed topics for summarization. The drug treatment citations were also processed with the conventional treatment point-of-view summarization software, using the same seed topics. We built reference standards of drug treatment and

preventive interventions by forwarding DynaMed recommendations to two reviewers, who highlighted interventions that they thought were credible. An adjudicator resolved disagreements between the two reviewers. The reviewers and the adjudicator were MDs.

I also built baselines by processing citation text with MetaMap to extract information that a clinician might find if directly reviewing PubMed output. The baseline methodology was based on techniques developed by Fiszman, [29] Zhang, [16]and their colleagues. Citation data were processed with MetaMap, retaining terms from desired semantic groups. Terms whose frequencies exceeded a threshold of the mean of all retained topic term frequencies, plus one standard deviation, formed the baseline for each disease topic/point-of-view pairing.

I compared summarization output to the reference standards. For output originating from the citations focused on drug treatment, I only looked at semantic predications in the form "Intervention X_TREATS_Disease_Y", where the object argument was an expression of the topic disease. If the subject argument was one of the reference standard interventions, that intervention received a true positive status. If the subject argument was potentially a general expression of a reference standard intervention, I examined the span of citation text which the predication captured. If the reference standard intervention was indicated in the span of text, it received a true positive status. I used the same methodology in evaluating the output for prevention, with one exception. Because the pilot project had confirmed that other types of predications in addition to those in the form of "Intervention X_PREVENTS_Disease_Y" could provide relevant data, I examined all predications originating from the PubMed queries addressing disease prevention.

I calculated recall, precision, and F-score using the same procedures used in the pilot project addressing disease prevention, with one exception regarding precision. All system output semantic predications were used in calculating precision for disease prevention and drug treatment, whereas in the pilot project only predications in the form "Intervention X_PREVENTS_Disease Y" were used in computing precision.

Results. The PubMed queries produced 19,422 citations focused on drug treatment for the four disease topics, and 1735 citations addressing prevention. SemRep produced a total of 162,184 semantic predications from the drug treatment citations, and 10,763 predications from the prevention citations. Dynamic summarization yielded a total of 20,616 semantic predications originating from the drug treatment citations for the four disease topics, and 811 citations originating from the prevention citations. The conventional software application produced 13,134 predications originating from the drug treatment citations.

Dynamic summarization drug treatment output produced 0.848 average recall and 0.377 average precision. The conventional application produced average recall and precision scores of 0.583 and 0.712 for drug treatment. The baseline methodology yielded average recall and precision scores of 0.234 and 0.306. For prevention output, dynamic summarization produced an average recall of 0.655 and an average precision rate of 0.329. The baseline produced an average recall of 0.269 and an average precision of 0.247.

The work for Aim 3 is described in detail, in Chapter 5. The following four chapters provided detailed descriptions of the methods, results, and implications of the work to complete each aim

References

1. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006 Feb;7(2):119-29.

2. Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J. 2006 Mar;23(1):3-12.

3. Chang AA, Heskett KM, Davidson TM. Searching the literature using medical subject headings versus text word with PubMed. Laryngoscope 2006 Feb;116(2):336-40.

4. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch T. Semantic MEDLINE: a web application to manage the results of PubMed searches. . Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, 2008: 69 -76.

5. Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc. 2010 Oct;98(4):273-81.

6. Workman TE, Hurdle JF. Dynamic summarization of bibliographic-based data. BMC Med Inform Decis Mak. 2011;11:6.

7. Workman TE, Stoddart JM. Rethinking information delivery: using a natural language processing application for point-of-care data discovery. J Med Libr Assoc. Forthcoming 2011.

8. Workman TE, Hurdle JF. Text summarization as a decision support aid. 2011.

9. Liddy ED. Natural Language Processing. In: Drake MA, ed. Encyclopedia of library and information science. New York, NY: Marcel Dekker, 2003.

10. Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

11. Turing AM. Computing machinery and intelligence. Mind 1950;59:433 - 60.

12. Chomsky N. Three models for the description of language. IRE Transactions on Information Theory 1956;2(3):113 - 24.

13. Luhn HP. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 1958;1(4):309 - 17.

14. Hahn U, Mani I. The challenges of automatic summarization. Computer 2000;33(11):29 - 36

15. Mani I, Maybury M, eds. Advances in automatic text summarization. Cambridge, MA: MIT Press, 1999.

16. Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform. 2011 May 8.

17. Mani I. Automatic summarization. Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2001.

18. Yang J, Cohen AM, Hersh W. Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts. AMIA Annu Symp Proc. 2007:831-5.

19. McKeown K, Elhadad N, Hatzivassiloglou V. Leveraging a common representation for personalized search and summarization in a medical digital library. 3rd ACM/IEEE-CS Joint Conference on Digital libraries. Houston, TX, 2003: 159–70.

20. Yoo I, Hu X, Song IY. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC Bioinformatics 2007;8 Suppl 9:S4.

21. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, Ely J, Yu H. AskHERMES: An online question answering system for complex clinical questions. J Biomed Inform. 2011 Apr;44(2):277-88.

22. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

23. Rindflesch T, Fiszman M, Kilicoglu H, Libbus B. Semantic knowledge representation project; a report to the Board of Scientific Counselors. Lister Hill National Center for Biomedical Communications, 2003.

24. Smith L, Rindflesch T, Wilbur WJ. MedPost: a part-of-speech tagger for bioMedical text. Bioinformatics 2004 Sep 22;20(14):2320-1.

25. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

26. Koldehoff M, Zakrzewski JL. Taurolidine is effective in the treatment of central venous catheter-related bloodstream infections in cancer patients. Int J Antimicrob Agents. 2004 Nov;24(5):491-5.

27. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, 2004:76-83.

28. Roberts PM. Mining literature for systems biology. Brief Bioinform. 2006 Dec;7(4):399-406.

29. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.

30. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc. 2006:254-8.

31. Sneiderman C, Demner-Fushman D, Fiszman M, Rosemblat G, Lang FM, Norwood D, Rindflesch TC. Semantic processing to enhance retrieval of diagnosis citations from Medline. AMIA Annu Symp Proc. 2006:1104.

32. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

33. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc. 2003:554-8.

34. Libbus B, Kilicoglu H, Rindflesch TC, Mork JG, Aronson A. Using natural language processing, locus link, and the gene ontology to compare OMIM to MEDLINE. Proceedings of the HLT-NAACL Workshop on Linking the Biological Literature, Ontologies and Databases: Tools for Users 2004: 69-76.

35. The Perl Programming Language: Perl.org. [cited 19 Sept 2011]. <http://www.perl.org>.

36. Mitchell JA, McCray AT. The Genetics Home Reference: a new NLM consumer health resource. AMIA Annu Symp Proc 2003:936.

37. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2002 Jan 1;30(1):52-5.

38. Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics 1951;22(1):79 – 86.

39. Riloff E. Automatically generating extraction patterns from untagged text. Proceedings of the Thirteenth National Conference on Artificial Intelligence. Menlo Park, CA: The AAAI Press/MIT Press, 1996: 1044–9.

40. Manning CD. Foundations of statistical natural language processing. Cambridge, MA: MIT Press, 1999.

41. Manning CD. Introduction to information retrieval. New York, NY: Cambridge University Press, 2008.

42. Mani I, Bloedorn E. Machine learning of generic and user-focused summarization. Proceedings Fifteenth National Conference on Artificial Intelligence (AAAI-98) Tenth Conference on Innovative Applications of Artificial Intelligence. Madison, WI, 1998: 821-6.'

43. Resnik P. Selectional constraints: an information-theoretic model and its computational realization. Cognition 1996;61:127-59.

44. DynaMed. Ipswitch, MA: EBSCO Publishing, 1995 - [1995 - cited 15 September 2011]. <http://www.ebscohost.com/DynaMed/>.

45. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998 Oct 21;280(15):1347-52.

46. Ely JW, Osheroff JA, Maviglia SM, Rosenbaum ME. Patient-care questions that physicians are unable to answer. J Am Med Inform Assoc. 2007 Jul-Aug;14(4):407-14.

CHAPTER 2

BIOMEDICAL TEXT SUMMARIZATION TO SUPPORT GENETIC DATABASE CURATION: USING SEMANTIC MEDLINE TO CREATE A SECONDARY DATABASE OF GENETIC INFORMATION

T. Elizabeth Workman, M.L.I.S., Marcelo Fiszman, M.D., Ph.D., John F. Hurdle, M.D., Ph.D., Thomas C. Rindflesch, Ph.D.

Journal of the Medical Library Association: JMLA 2010 Oct;98(4):273-81

Abstract

Objective: This paper examines the development and evaluation of an automatic summarization system in the domain of molecular genetics. The system is a potential component of an advanced biomedical information management application called Semantic MEDLINE and could assist librarians in developing secondary databases of genetic information extracted from the primary literature . **Methods:** An existing summarization system was modified for identifying biomedical text relevant to the genetic etiology of disease. The summarization system was evaluated on the task of identifying data describing genes associated with bladder cancer in MEDLINE citations. A gold standard was produced using Genetics Home Reference and Online Mendelian Inheritance in Man (OMIM) records. Genes in text found by the system were compared to the gold standard; recall, precision, and F-measure were calculated. **Results**: The system achieved recall of 46%, and precision of 88% (F-measure = 0.61) by taking GeneRIFs into account. **Conclusion:** The new summarization schema for genetic etiology has potential as a component in Semantic MEDLINE to support the work of data curators.

Introduction

Due to evolving technologies and policies, libraries have an increasing interest in the process of data curation. As McDonald and Uribe point out [1], the open access movement, coupled with ever-increasing volumes of data from current scientific investigations, has created a research environment which calls for new management strategies for domain-specific data curation, which is defined here as the organization,

preservation, and enhancement of the data through value-added features such as annotations . This environment has united traditional academic participants such as librarians, researchers, and administrators, who previously worked independently.

Librarians have the opportunity to take a leadership role in implementing techniques and policies for data curation and preservation. For example, an academic library could partner with other campus departments in creating the framework for enhancing and preserving the institution's research, possibly creating unique and priceless resources. There are several examples in which librarians have taken the lead in information curation, access, preservation, and management, in neuro-ophthalmology [2], institutional repositories [3], and other areas. Curators of secondary databases face the demanding task of identifying relevant information from primary sources, which are continually increasing [4]. The development of curated databases is often based on a complex methodology of information discovery, content development, and expert review [5] [6].

Information discovery for secondary databases may be dependent on traditional information retrieval and the meticulous, manual inspection of documents resulting from conventional searches of databases such as MEDLINE. This task can be quite daunting and time consuming. In developing the Human Protein Reference Database (HPRD), for example, developers performed extensive searches in PubMed to identify relevant literature. Then, researchers spent over 50,000 hours during an eight-month period reading more than 300,000 articles to manually curate HPRD records [7].

Biomedical information retrieval techniques provide support for secondary database curation [8]; however, little research has been published on using automatic summarization to augment these techniques and help manage the information contained

in the large numbers of MEDLINE citations often returned by PubMed searches. Automatic summarization provides the information most relevant to a user's interest from a source in a condensed format. The advanced biomedical information management application Semantic MEDLINE [9] (public demonstration interface at http://skr3.nlm.nih.gov/SemMedDemo/) integrates automatic summarization with information retrieval, semantic processing, and visualization to analyze biomedical text. Semantic processing in the application uses SemRep [10] [11] to represent document content as semantic relations (e.g. Drug X TREATS Disease Y), also referred to as semantic predications. Automatic summarization [12] further processes these relations to identify those that are most relevant to a user's needs. The resulting semantic relations are then presented to the user in a graph that visually displays the content of retrieved documents. Since links are maintained between semantic relations and input text, the graph serves as a guide to help users decide what to read.

The thrust of the research reported here was to extend the use of Semantic MEDLINE to the domain of molecular genetics. Librarians maintaining databases in this domain must keep pace with the growing amounts of data generated by improved genetic analytic technologies [13] and need the ability to easily identify genes associated with a particular disease. The authors first describe the technology required to extend Semantic MEDLINE and then suggest how the application can serve as an adjunct to traditional information retrieval in secondary database curation. In the evaluation, genes extracted by the system were compared to those found in two actively curated genetic databases, Genetics Home Reference and Online Mendelian Inheritance in Man.

Background

Curated Resources

Genetics Home Reference [14], hosted by the National Library of Medicine, was introduced in 2003 as a consumer-friendly Website for genetic diseases [15]. The site implements a content development strategy that combines human effort with select complementary automated functions [16]. The Online Mendelian Inheritance in Man (OMIM) database [17], a Johns Hopkins University product hosted by The National Center for Biotechnology Information at the National Library of Medicine, implements a curation strategy in which journal content is daily reviewed by hand [18] [19]. Under agreement with publishers, OMIM receives articles from specific journals prior to publication. OMIM staff also read additional publications looking for potential materials for manual review. Genetics Home Reference provides information on a level appropriate for patients; OMIM furnishes more technical, detailed genetic disease information that is very suited for scientists. The two databases provide a full landscape of online genetics information.

Document Source

The primary document source for this study was MEDLINE, the premier database of the National Library of Medicine, which includes over 18 million citations, representing the biomedical literature from 1949 to the present [20].

Semantic MEDLINE

Semantic MEDLINE [9] is a multiple-step tool in development that helps users manage the results of PubMed searches. The application extracts the succinct meaning of the text it processes and displays the resulting distilled data in an interactive graph that maintains links to the original text. Semantic MEDLINE proceeds in four steps: PubMed searching, extraction of semantic predications with SemRep, automatic summarization, and visualization (Figure 3).

SemRep

At the core of Semantic MEDLINE is SemRep [10][11], a rule-based, symbolic natural language processing application that uses the Unified Medical Language System (UMLS) [21] to express the meaning of text in a straight-forward and consistent representation, called a semantic predication. Such a representation has arguments and a predicate. The following illustrates this process:

Original text:

> "The *IGF1R* is up-regulated *in bladder cancer* compared with non-malignant bladder, and might contribute to a propensity for invasion [22]."

Extracted semantic predication:

> IGF1R gene ASSOCIATED_WITH Carcinoma of bladder

SemRep uses MetaMap [23] to map the text *IGF1R* and *bladder cancer* to the Metathesaurus concepts "IGF1R Gene" and "Carcinoma of Bladder," which are associated with *semantic types* (or classes) 'Gene or Genome' and 'Neoplastic Process', respectively. These concepts function as the arguments of the predication. Based on the

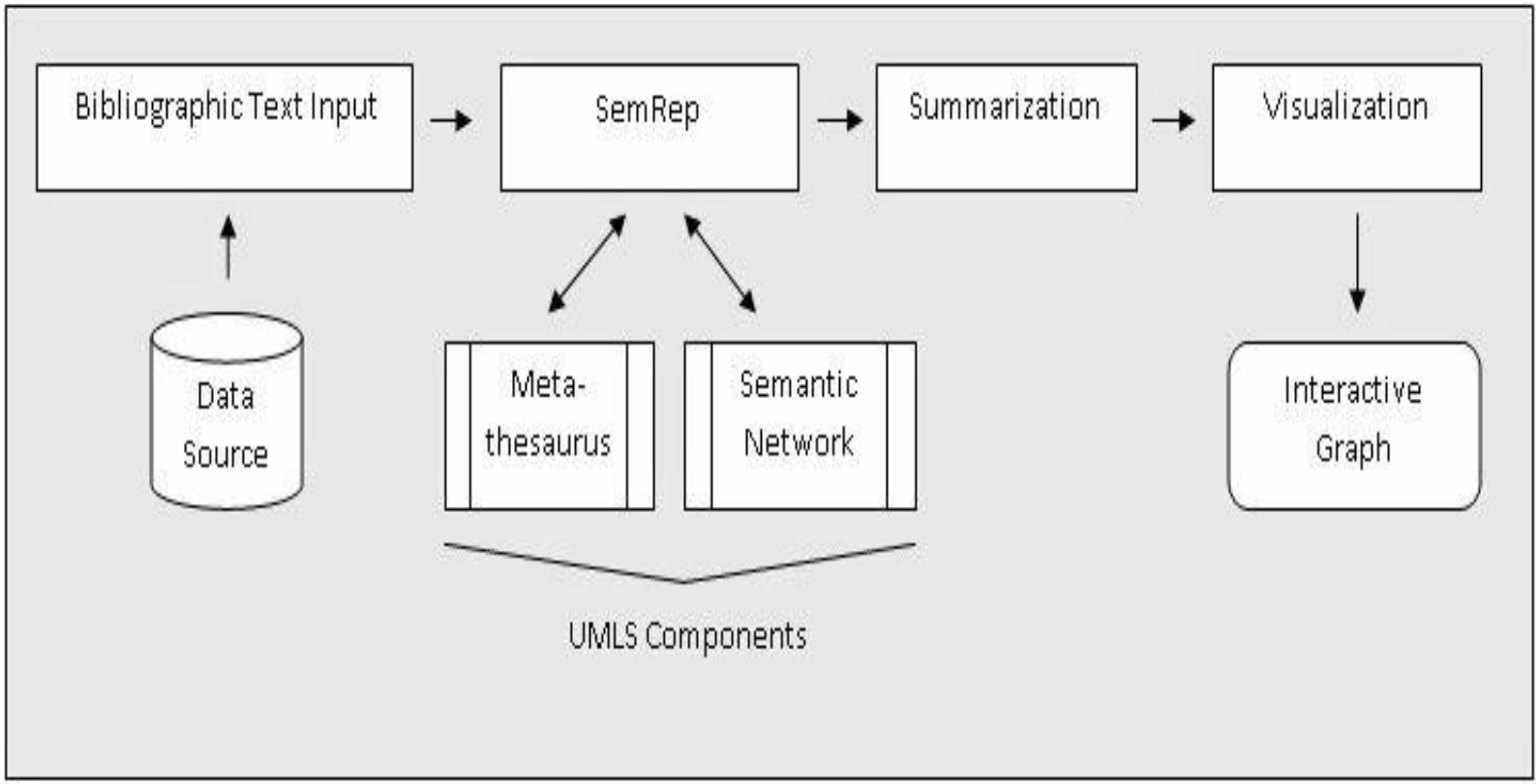

Figure 3: Semantic MEDLINE

semantic types, SemRep then draws upon the Semantic Network to identify the *predicate* (or relation), ASSOCIATED_WITH, that binds these arguments. SemRep extracts semantic predications for an array of predicates, including TREATS, LOCATION_OF, INHIBITS, INTERACTS_WITH, CAUSES, PREDISPOSES, and ASSOCIATED_WITH, among others.

Automatic Summarization

In the summarization phase, a schema filters semantic predications extracted from MEDLINE citations according to a user-selected *point-of-view* and *topic concept* [12]. For example, if a user were interested only in information addressing treatment (i.e., the point of view) for a particular disease (i.e., the topic concept), summarization would collect the best predications that expressed this information. The summarization architecture does this by subjecting SemRep predications to four sequential phases of filtering, which select only those semantic predications pertinent to the selected point of view and topic concept:

Relevance: collects predications addressing the user-selected topic concept.

Connectivity: augments relevancy predications with others associated with the topic concept.

Novelty: eliminates predications asserting basic knowledge that users already know.

Saliency: limits final output to predications that occur most frequently.

The current online Semantic MEDLINE prototype includes schemas that summarize for treatment [12], substance interactions [24], diagnosis, and pharmacogenomics [25] points of view.

Methods

In order to explore Semantic MEDLINE's ability to assist librarians in curating secondary genetics databases, a new summarization schema was first created, targeting semantic predications that are relevant to the genetic etiology of disease. Subsequently, documents retrieved from MEDLINE were processed within the Semantic MEDLINE model enhanced with this schema. Finally, the genes identified during this processing were evaluated by comparing them to a reference standard compiled from Genetics Home Reference and OMIM.

A Summarization Schema for Genetic Etiology of Disease

As noted earlier, a schema provides a general means of identifying SemRep predications for a particular point of view. Earlier work [26, 27] had enhanced SemRep to extract semantic predications on the genetic etiology of disease, but had not provided a summarization schema. A schema for this purpose has two features: a list of allowable predicates, and a list of semantic types which specify which Metathesaurus concepts the listed predicates are permitted to have as arguments. The new schema was designed in such a way as to summarize SemRep data for any disease topic the user may choose, from the point of view of genetic disease etiology.

In crafting the schema, allowable semantic types were assembled into three groups: genetic phenomenon, anatomy, and disease process. The following indicates the UMLS semantic types included in each of these groups:

Genetic phenomenon: Amino Acid Sequence; Enzyme; Genetic Function; Nucleic Acid, Nucleoside, or Nucleotide; Nucleotide Sequence; Amino Acid, Peptide, or Protein; Gene or Genome; and Molecular Sequence.

Anatomy: Anatomical Structure; Body Part, Organ, or Organ Component; Cell; Cell Component; Embryonic Structure; Fully Formed Anatomical Structure; Gene or Genome; and Tissue.

Disease Process: Acquired Abnormality; Anatomical Abnormality; Congenital Abnormality; Cell or Molecular Dysfunction; Disease or Syndrome; Injury or Poisoning; Mental or Behavioral Dysfunction; Neoplastic Process; Pathologic Function; Sign or Symptom; Biologic Function; Cell Function; Mental Process; Molecular Function; Natural Phenomenon or Process; Organism Function; Organ or Tissue Function; Physiologic Function; Behavior; Mental or Behavioral Dysfunction; and Finding.

The schema for genetic etiology of disease allows the following predicates: AFFECTS, ASSOCIATED_WITH, AUGMENTS, CAUSES, DISRUPTS, COEXISTS_WITH, INHIBITS, PREDISPOSES, and STIMULATES. When the arguments of these predicates are limited to the semantic types noted above, the schema specifies the semantic predications permitted in summarization when generated from the point of view of the genetic etiology of disease. The following illustrates the specific semantic types (by the previously noted groups) and predicate combinations allowed by the schema:

{genetic phenomenon} AFFECTS {disease process}

{genetic phenomenon} AUGMENTS {disease process}

{genetic phenomenon} DISRUPTS {disease processes OR anatomy}

{genetic phenomenon} ASSOCIATED_WITH {disease process}

{genetic phenomenon} PREDISPOSES {disease process}

{genetic phenomenon} CAUSES {disease process}

{genetic phenomenon} STIMULATES {genetic phenomenon}

{genetic phenomenon} INHIBITS {genetic phenomenon}

{disease process} COEXISTS_WITH {disease process}

For example, this schema allows the genetic etiology predication “NAT 2 gene PREDISPOSES Carcinoma of Bladder” to be included in the summary because the predicate PREDISPOSES matches, and further, the subject argument “NAT 2 gene” has semantic type ‘Gene or Genome’, which is included in the “genetic phenomenon” group and the object argument has semantic type ‘Neoplastic Process’, which is in the “disease process” group. The use of three semantic groups permits predications in the summary that do not strictly assert genetic etiology, but rather provide likely valuable additional information, such as “{genetic phenomenon} DISRUPTS {anatomy}” and “{disease process} COEXISTS_WITH {disease process}.” Finally, the predication “Immunotherapy TREATS Carcinoma of Bladder” is not allowed, because the predicate TREATS is not in the schema.

Input Text Acquisition

In order to test the efficiency of the Semantic MEDLINE model (enhanced with the new schema) in identifying research literature relevant to curation of a secondary resource, the team chose bladder cancer, the sixth overall leading form of cancer in the

U.S. [28], as a topic of study. To complete the first phase in the Semantic MEDLINE model, the project team executed the following PubMed query:

urinary bladder neoplasms[mh] OR "bladder cancer" OR "cancer of the bladder"

Limits: Publication Date from 2003/01/01 to 2008/07/31, only items with abstracts, English

Five thousand, six hundred and six citations (titles and abstracts) were retrieved with this query and subsequently downloaded from MEDLINE.

Document Processing

All citations were processed by SemRep and the extracted predications were then submitted to the new schema for summarization on the topic of bladder cancer according to the genetic etiology of disease point of view.

Extracting a List of Genes from the Summarized Predications

A list of genes implicated in bladder cancer was extracted from the predications in the summarization schema's output, subject to the following criteria: 1) the subject concept must have a semantic type belonging to the group "genetic phenomenon" and the object must be a concept referring to bladder cancer ("Carcinoma of bladder," "Bladder Neoplasm," and "Carcinoma, Transitional Cell"). These bladder cancer concepts map to the semantic type "Neoplastic Process," which is in the "disease process" group. For example, "FGFR3 gene" is extracted from the "FGFR3 gene ASSOCIATED_WITH Carcinoma, Transitional Cell."

Compiling the Reference Standard from OMIM and Genetics Home Reference

The reference standard for this project consisted of the genes noted as associated with bladder cancer in OMIM and Genetics Home Reference. In order to identify valid genes in OMIM, we retrieved all records which were either phenotypically relevant to bladder cancer, or which provided clinical synopses for this disease, using the following query: "bladder cancer"[All Fields] OR "bladder cancers"[All Fields] OR "bladder cancer cases"[All Fields] OR "bladder cancer cell"[All Fields] OR "bladder cancer patients"[All Fields] OR "bladder carcinoma"[All Fields] OR "bladder carcinogenesis"[All Fields]

This query was first executed with the OMIM interface limits options manipulated to retrieve a broad range of genetic information associated with bladder cancer, varying from known genes with known chromosome loci, hypothesized loci only, to a suspected, but not ascertained genetic basis. Then, the query was issued a second time after modifying the OMIM interface limits options to retrieve only records which included a clinical synopsis. The results of these two queries were then combined, resulting in 14 records. In Genetics Home Reference, the query “bladder” retrieved records either addressing general phenotype information (with the general label “Genetic Condition”) or a gene. Of these, we identified 11 records containing information relevant to the genetic basis of bladder cancer.

The 25 records extracted from OMIM and Genetics Home Reference were then examined for specific genes. Records were limited to those based on source literature published within the study’s timeframe (January 2003 – July 2008). Genetics Home Reference records noted 10 genes with disease implications, while OMIM noted seven;

four genes were noted by both databases as relevant to bladder cancer. Genes noted in each record were classified as having a *confirmed or possible* involvement in bladder cancer. Genes noted in the main phenotype records of each database as implicated in bladder cancer were classified as having a *confirmed* involvement. To illustrate, the FGFR3 gene received a *confirmed* classification, due to its combination with the phrase "implicated in bladder carcinogenesis" in OMIM record #109800 for bladder cancer [29], and for its presence in the Genetics Home Reference bladder cancer condition record, indicating that it is "associated with bladder cancer" [30]. Genes noted in other records in certain explicit contexts (adjacent to survival rates, for example) received a *possible* classification. For example, Genetics Home Reference notes an "amplification" of the *possible*-classified ERBB3 gene "and/or overexpression of [its] protein" in bladder tumors in the ERBB3 gene record [31]. Genes tied to conflicting, uncertain, or undefined wording were also classified as *possible*. For example, Genetics Home Reference notes conflicting evidence defining the ATM gene's implication in bladder cancer [32]. Therefore, it was assigned a *possible* classification. All genes from GHR and OMIM, regardless of classification, were included in the final reference standard as implicated in bladder cancer. Using these criteria, 13 genes were included in the reference standard (Table 1).

Evaluation

The second author (MF) manually matched the output of the genes extracted from the final summarization output against the genes in the reference standard. Based on this matching, recall, precision, and F-measure were calculated. Recall was defined as the

Table 1 – Gold standard genes associated with bladder cancer.

Gene Symbol	Source	Classification
FGFR3	Both	Confirmed
XPD	OMIM	Confirmed
RAG1	OMIM	Confirmed
TP53	Both	Confirmed
MTCYB	OMIM	Confirmed
HRAS	Both	Confirmed
NAT2	Both	OMIM Confirmed; GHR Possible
RB1	GHR	Confirmed
TSC1	GHR	Confirmed
ATM	GHR	Possible
TGFB1	GHR	Possible
MDM2	GHR	Possible
ERBB3	GHR	Possible

percentage of genes in the reference standard which were found in the summarized output. Precision was measured by determining the percentage of all genes in the summarized output that was noted in the reference standard, or in an Entrez Gene [33] GeneRIF, as implicated in bladder cancer development. Gene References into Function (GeneRIF) annotations [34] in corresponding Entrez Gene records (for *Homo sapiens* only) were consulted for such genes which were not noted in OMIM or Genetics Home Reference. If an explicit GeneRIF annotation noted an association of the gene with bladder cancer, it was counted as a true positive in the precision computation. The F-measure, which ranges from a high of 1 to a low of 0, expresses a balanced average between the recall and precision scores.

Results

Predications and Genes Extracted

SemRep extracted 38,498 semantic predications from the 5606 citations retrieved from MEDLINE. The summarization phase limited these to 359 semantic predications relevant to bladder cancer (using the schema for genetic etiology). From these predications 17 genes and proteins were extracted based on the criteria noted in section 3.4. These were normalized to the gene name in Entrez Gene and are shown in Table 2.

Table 3 shows the results of manually comparing the genes from summarization to the reference standard (OMIM and Genetics Home Reference) to compute recall, and to Entrez Gene GeneRIFs in addition to the reference standard for computing precision. Of the 13 genes in the reference standard, six were represented in the final summarization output. Out of 17 genes in the summarization output, 11 were false positives when compared only to the reference standard, while only two were false positives when compared to the reference standard and GeneRIFs.

Discussion

The modified summarization system described in this paper and evaluated with bladder carcinoma genes obtained moderately good recall when compared to the reference standard compiled from OMIM and Genetics Home Reference. Precision increased substantially when GeneRIFs were taken into account. GeneRIF annotations are routinely added to an Entrez Gene record when the linked PubMed record is indexed, as part of an indexer's work, and can provide additional insight into a gene's involvement in a disease process.

Table 2. Genes extracted by the summarization program.

Summarization Output
TP53 gene
FGFR3 gene*
BIRC5 gene
Cadherins (CDh1)**
cyclooxygenase 2 (PTGS2)
CDKN2A gene
CDC91L1 gene
Candidate Disease Gene
NAT2 gene
EGF gene
TGFB1 protein, human (TGFB1)
MDM2 gene
HRAS gene
GSTT1 gene
GSTM1 gene
Gelatinase B (MMP9)
CD82 gene

**Genes that appear in the reference standard associated with bladder cancer are in bold.
*Genes normalized from proteins are presented in parentheses

Table 3. Performance measures* for the summarization system on extracting genes related to bladder cancer from MEDLINE.

Metric	
Precision	88%
Recall	46%
F-measure	0.61

* The table displays the results with taking GeneRIFs into account for assessing precision (as explained in Methods).

There are two reasons for the level of current results. SemRep processing contributed to some errors, and further development to improve the accuracy of this application is part of ongoing research. In addition, genes are noted as implicated in a disease process in OMIM and Genetics Home Reference due to curation decisions which are in part independent of what is noted in the collective professional literature (and hence in SemRep output). GeneRIFs, on the other hand, are routinely created as part of the indexing process for all MEDLINE citations which include gene information. For example, the "CDC91L1 gene" was commonly noted as related to bladder cancer in the summarized SemRep output, but was not noted in the OMIM and Genetics Home Reference records consulted in creating the reference standard, eventhough one of the GeneRIFs in Entrez Gene for CDC91L1 in *homo sapiens* notes the following: "CDC91L1 (PIG-U) is a newly discovered oncogene in human bladder cancer" (PMID – 15034568, published within the time frame of this study). In an actual application, summarized output could guide curation, but it would be up to curators to decide what information would be included in their secondary databases.

The Semantic MEDLINE process, implementing SemRep, summarization, and visualization, converts large amounts of data into a concise representation of semantic predications expressing the data's meaning, which can then be quickly reviewed and traced back to the original text. This process can potentially save time for database curators reviewing large amounts of information (although our project did not test this hypothesis).

Using the modified schema presented in this paper, the genetic summary can be displayed in Semantic MEDLINE as an interactive graph [9] (See Figure 4). Arcs (the

lines connecting the labeled concepts) represent relations between each argument node (the labeled concepts). The central node in the graph represents the user-determined topic of the summary ("Carcinoma of bladder"). The user may select or deselect predicates in the upper-right side panel, to focus on specific relationships in the graph. By right-clicking on a given arc, the user can access the original text from which a semantic predication was extracted. In Figure 4, the user may right-click the "PREDISPOSES" relationship arc between the GSTT1 gene concept node and the central concept "Carcinoma of bladder" to view the original text (a MEDLINE citation).

As noted in the introduction, use of this tool creates the potential for collaborative curation work between librarians and researchers. The following scenario further illustrates how this might work in practice: The board that oversees the institutional repository at a major university decides to integrate into this repository primary data from a university laboratory exploring the genetic etiology of disease. The librarian in charge of repository curation notes that an added-value resource summarizing the published findings of the laboratory's research would assist other campus scientists to appraise the data. The librarian submits a query to Semantic MEDLINE to locate and download all relevant citations published by the laboratory's faculty. The librarian then uses the application to sequentially summarize the MEDLINE data for each disease studied, from the point of view of genetic etiology. To review the summarized results, the librarian visualizes the data for each disease, clicking on the arcs within the graph to view citations associated with each semantic predication. Using the summarized data, the librarian creates a concise report of the findings associated with the lab's data. The report is stored in the institutional repository with the lab's research data, so that users can quickly

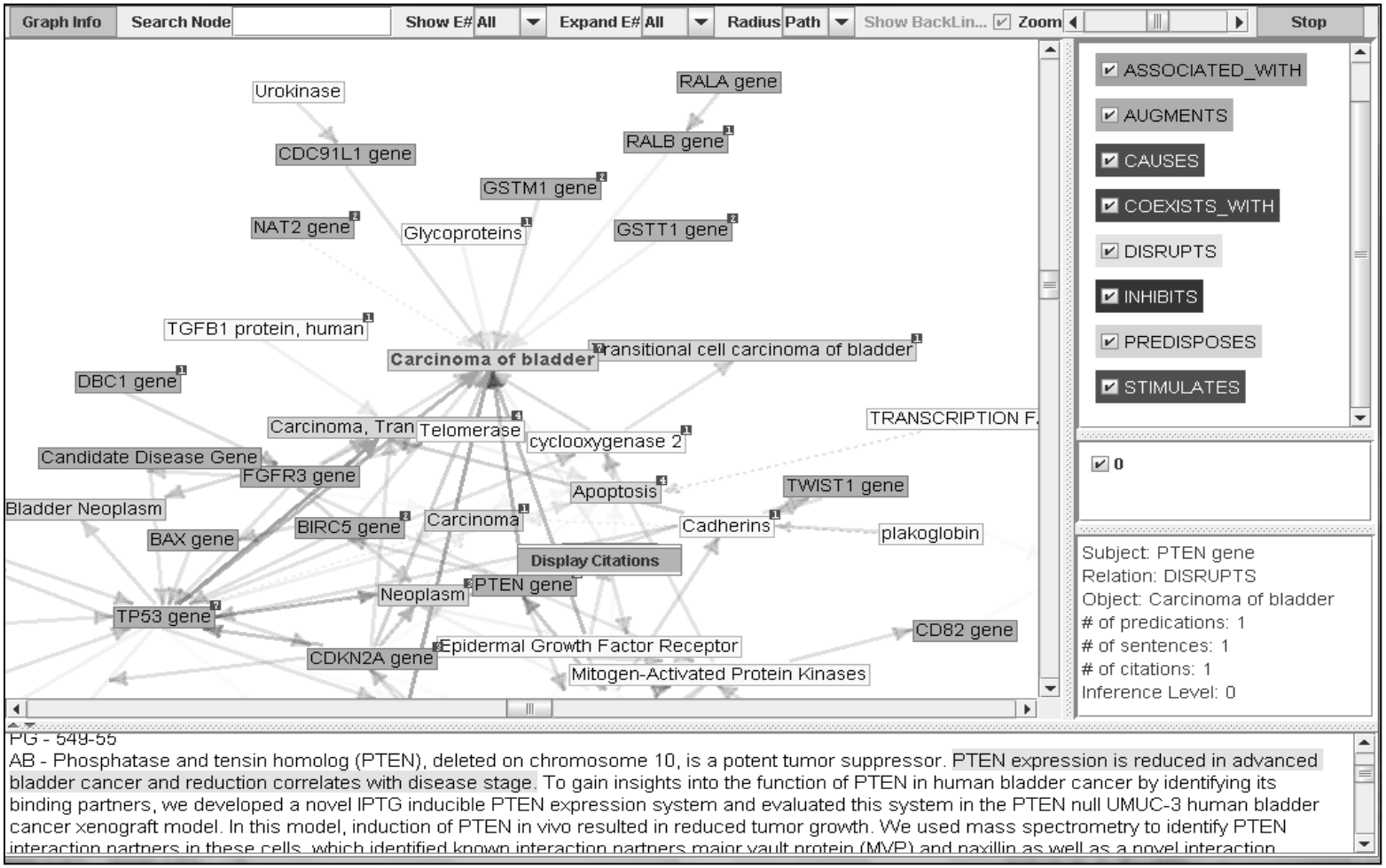

Figure 4: Visualization graph illustrating summarized semantic predications

determine its potential relevance in their own endeavors.

Limitations of the Study

The evaluation was performed with one disease and it is hard to predict the generalizability of performance when more diseases are taken into account. However, SemRep and the summarization system components of Semantic MEDLINE have been proven to be effective in a topic-oriented evaluation study to support evidence-based medical treatment of 50 diseases [35]. Performance will likely scale similarly to potentially support genetic database curation. A further limitation is that the natural language processing system (SemRep) does not have access to information curators use to decide what genes are established markers for diseases. These are curation policies that go beyond any language processing system.

Conclusions

Semantic MEDLINE transforms vast amounts of bibliographic text into succinct, brief statements. To place this in a quantitative perspective, in this study Semantic MEDLINE reduced 5606 MEDLINE citations to 359 semantic predications. Curators could substantially reduce the amount of time needed to manually review original MEDLINE documentation by first processing it with Semantic MEDLINE and then reviewing its output.

This study explored the application of Semantic MEDLINE for a specific task, that of database curation. As noted before, this task is relevant to emerging opportunities for librarians to contribute to parent organizations and the scientific community at large, as

professional partners. Other work may also be aided by Semantic MEDLINE applications. For example, librarians could assist patrons in quickly assessing large amounts of bibliographic text by first processing it with Semantic MEDLINE, and then instructing them on using its interactive visual display. Outcomes from separate groups of research studies, represented as bibliographic text, could be compared. These services could reaffirm the importance of university library services, and strengthen the role of librarians as essential partners in the research endeavors of their individual institutions.

Future work in schema development and domain exploration is needed in order to extend Semantic MEDLINE's capabilities and to measure its effectiveness. Summarization which accommodates points of view beyond those currently available will enable the system to process data for additional needs. Assessing Semantic MEDLINE's ability to assist in additional tasks such as point of care information delivery and patient education will give further insight to its potential uses.

Acknowledgements

The authors express their gratitude to Graciela Rosemblat for assisting with the study's evaluation, to Jeanne Le Ber for editorial assistance, and to Joyce Mitchell for advice and suggestions. They also wish to thank the National Library of Medicine for funding this project through grant number T15LM007123 and other program funding, and the Oak Ridge Institute for Science and Education for administering part of the funding.

References

1. MacDonald S, Uribe LM. Libraries in the converging worlds of open data, e-research, and Web 2.0. Online 2008 Mar/April 32(2): 36-40.

2. Digre KB, Lombardo NT, Frohman L. Neuro-ophthalmology virtual education library (NOVEL). Neuro-Ophthalmology Oct 2007; 31(5): 175-8.

3. Koopman A, Kipnis D. Feeding the fledgling repository: starting an institutional repository at an academic health sciences library. Med Ref Serv Q. 2009 Summer; 28(2):111-22.

4. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews Genetics 2006. Feb. 7: 119-29.

5. Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J. DATF: a database of Arabidopsis transcription factors. Bioinformatics 2005 May 15; 21(10):2568-2569.

6. Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J. DRTF: a database of rice transcription factors. Bioinformatics 2006 May 15; 22(10):1286-1287.

7. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003 Oct; 13(10):2363-71.

8. Caspi R, Fulcher C, Ingraham J, Keseler I, Krummenacker M, Paley S. Curator's guide for pathway/genome databases [Internet] Menlo Park, CA: SRI International July 2007 [Cited 1 July 2010]. <bioinformatics.ai.sri.com/ptools/curatorsguide.pdf>.

9. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC. Semantic MEDLINE: a web application for managing the results of PubMed searches Proceedings of the Third International Symposium for Semantic Mining in Biomedicine 2008, pp. 69-76.

10. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In Chen, Fuller, Hersh, and Friedman, eds. Medical informatics: knowledge management and data mining in biomedicine. Springer, 2005, 399-422.

11. 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 Dec;36(6):462-77.

12. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics. 2004:76-83.

13. Roberts PM. Mining literature for systems biology. Briefings in Bioinformatics 2006 Oct 7(4): 399 – 406.

14. National Library of Medicine. Genetics Home Reference. [Internet] Bethesda, MD: National Library of Medicine; 2003 [updated 2004 Apr 7; cited 17 Dec 2009]. <http://ghr.nlm.nih.gov/>.

15. Mitchell JA, McCray AT. The Genetics Home Reference: a new NLM consumer health resource. AMIA Annu Symp Proc. 2003:936.

16. Mitchell JA, Fun J, McCray AT. Design of Genetics Home Reference: a new NLM consumer health resource. JAMIA 2004 Dec 11(6): 439 – 447.

17. Online Mendelian Inheritance in Man, OMIM [Internet] Baltimore, MD: McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, and National Center for Biotechnology Information, National Library of Medicine; 1995 [cited 17 Dec 2009] <http://www.ncbi.nlm.nih.gov/omim/>.

18. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009 Jan;37(Database issue):D793-6. Epub 2008 Oct 8.

19. Letovsky SI, editor. Bioinformatics: databases and systems. Boston, MA: Kluwer Academic Publishers; 1999.

20. National Library of Medicine. Data, news and update information: PubMed update.[Internet]. Bethesda, MD: National Library of Medicine; 2001 [updated 19 April 2010; cited 20 April 2010].
<http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update>.

21. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

22. Rochester MA, Patel N, Turney BW, Davies DR, Roberts IS, Crew J, Protheroe A, Macaulay VM. The type 1 insulin-like growth factor receptor is over-expressed in bladder cancer. BJU Int. 2007 Dec;100(6):1396-401.

23. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus:

the MetaMap program. Proc AMIA Symp. 2001:17-21.

24. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc. 2006:254-8.

25. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

26. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc. 2003:554-8.

27. Libbus B, Kilicoglu H, Rindflesch TC, Mork JG, Aronson AR. Using natural language processing, Locus Link, and the Gene Ontology to compare OMIM to MEDLINE. Proceedings of the HLT-NAACL Workshop on Linking the Biological Literature, Ontologies and Databases: Tools for Users. 2004: 69-76.

28. U.S. Cancer Statistics Working Group. United States cancer statistics: 1999–2005 *incidence and mortality Web-based report[Internet]*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; 2009 [cited 18 Dec 2009]. <www.cdc.gov/uscs>.

29. #109800 Bladder cancer [Internet]. Baltimore, MD: Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, and National Center for Biotechnology Information, National Library of Medicine; 1995 [cited 18 Dec 2009]. <http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=109800>.

30. Bladder Cancer [Internet]. Bethesda, MD: Genetics Home Reference. National Library of Medicine; 2007 [published 13 Dec 2009; cited 18 Dec 2009]. <http://ghr.nlm.nih.gov/condition=bladdercancer>.

31. EERB3 [Internet]. Bethesda, MD: Genetics Home Reference. National Library of Medicine; 2009 [published 13 Dec 2009; cited 18 Dec 2009]. <http://ghr.nlm.nih.gov/gene=erbb3>.

32. ATM [Internet]. Bethesda, MD: Genetics Home Reference. National Library of Medicine; 2008 [published 13 Dec 2009; cited 18 Dec 2009]. <http://ghr.nlm.nih.gov/gene=atm>.

33. National Library of Medicine. Entrez Gene. [Internet] Bethesda, MD: National Library of Medicine; 2004 [cited 18 Dec 2009]. <http://www.ncbi.nlm.nih.gov/gene/>.

34. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc. 2003:460-4.

35. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct; 42(5):801-13.

CHAPTER 3

DYNAMIC SUMMARIZATION OF BIBLIOGRAPHIC-BASED DATA

T. Elizabeth Workman, M.L.I.S, John F. Hurdle, M.D., Ph.D.

BMC Medical Informatics and Decision Making 2011;11:6

Abstract

Background: Traditional information retrieval techniques typically return excessive output when directed at large bibliographic databases. Natural Language Processing applications strive to extract salient content from the excessive data. Semantic MEDLINE, a National Library of Medicine (NLM) natural language processing application, highlights relevant information in PubMed data. However, Semantic MEDLINE implements manually coded schemas, accommodating few information needs. Currently, there are only five such schemas, while many more would be needed to realistically accommodate all potential users. The aim of this project was to develop and evaluate a statistical algorithm that automatically identifies relevant bibliographic data; the new algorithm could be incorporated into a dynamic schema to accommodate various information needs in Semantic MEDLINE, and eliminate the need for multiple schemas. **Methods**: We developed a flexible algorithm named Combo that combines three statistical metrics, the Kullback-Leibler Divergence (KLD), Riloff's RlogF metric (RlogF), and a new metric called PredScal, to automatically identify salient data in bibliographic text. We downloaded citations from a PubMed search query addressing the genetic etiology of bladder cancer. The citations were processed with SemRep, an NLM rule-based application that produces semantic predications. SemRep output was processed by Combo, in addition to the standard Semantic MEDLINE genetics schema and independently by the two individual KLD and RlogF metrics. We evaluated each summarization method using an existing reference standard within the task-based context of genetic database curation. **Results**: Combo asserted 74 genetic entities implicated in bladder cancer development, whereas the traditional schema asserted 10 genetic entities;

the KLD and RlogF metrics individually asserted 77 and 69 genetic entities, respectively. Combo achieved 61% recall and 81% precision, with an F-score of 0.69. The traditional schema achieved 23% recall and 100% precision, with an F-score of 0.37. The KLD metric achieved 61% recall, 70% precision, with an F-score of 0.65. The RlogF metric achieved 61% recall, 72% precision, with an F-score of 0.66. **Conclusions**: Semantic MEDLINE summarization using the new Combo algorithm outperformed a conventional summarization schema in a genetic database curation task. It potentially could streamline information acquisition for other needs without having to hand-build multiple saliency schemas.

Background

The continued growth of bibliographic databases creates challenges to users practicing traditional information retrieval (IR) techniques. Standard search techniques, when applied to large databases such as PubMed, often return large, unmanageable lists of citations that do not fulfill the searcher's information needs [1, 2]. This problematic issue impedes many tasks, including secondary genetic database development. Databases such as Online Mendelian Inheritance in Man (OMIM) and Genetics Home Reference (GHR) use information from the biomedical literature to develop narrative records describing gene involvement in disease processes. Developers of secondary genetic databases built using the professional literature often rely on IR, and must invest much time and effort in procuring information [3]. The same problem prevents individuals from using IR effectively in other biomedical applications such as clinical decision support, [4] systematic review development, [5, 6] and even in Google searches [7].

NLP and Semantic MEDLINE

Natural language processing (NLP) can address this problem by identifying and summarizing text that fulfills a user's information needs in IR-procurable data. Examples of this approach include document clustering, [8] outcome polarity features in machine learning, [9] and content modeling in sentence selection [10]. NLP models leveraging transformations known as semantic predications can also address this issue. Semantic MEDLINE [11] is a multistage NLP system designed by researchers at the National Library of Medicine (NLM) to extract meaningful information from MEDLINE citations in the form of semantic predications, which are succinct declarations capturing the meaning of the original text. Its three core processes (Figure 5), SemRep, Summarization, and Visualization, respectively extracts semantic predications capturing the citations' content, identifies predications which are salient to a specific user-indicated information need, and displays them in a graphic representation (Figure 6). Currently, Semantic MEDLINE accommodates just a small handful of information needs, due to limitations in the Summarization stage. This problem renders Semantic MEDLINE to be an impractical tool for most users. We began this work intending to create an algorithm that would enable Semantic MEDLINE's Summarization stage to accommodate many information needs. To aid the reader in conceptualizing Semantic MEDLINE and our work to improve Summarization, we provide the following detailed description.

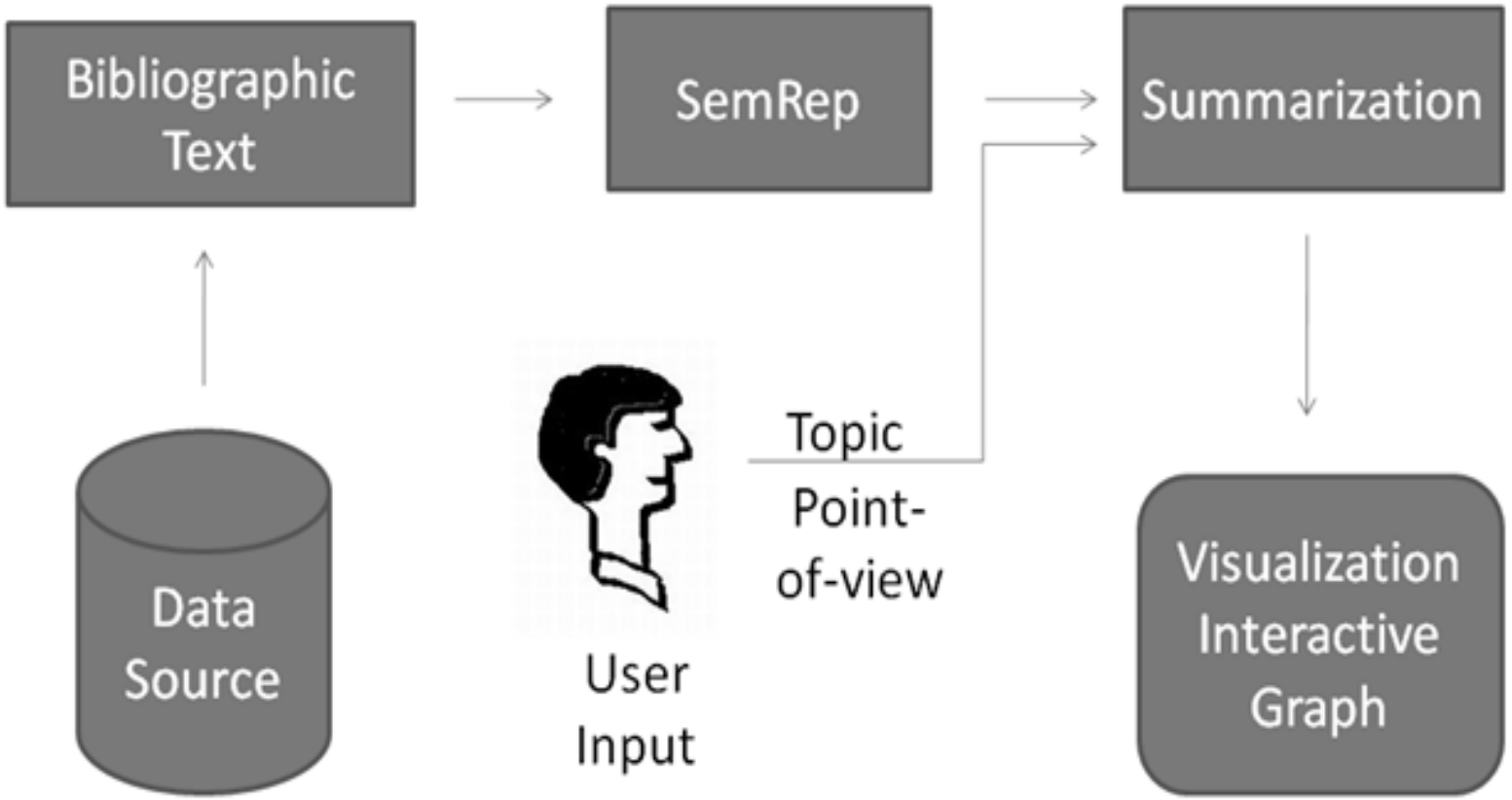

Figure 5. Semantic MEDLINE. The adaptive Combo algorithm described in this paper was designed to be incorporated into the Summarization process.

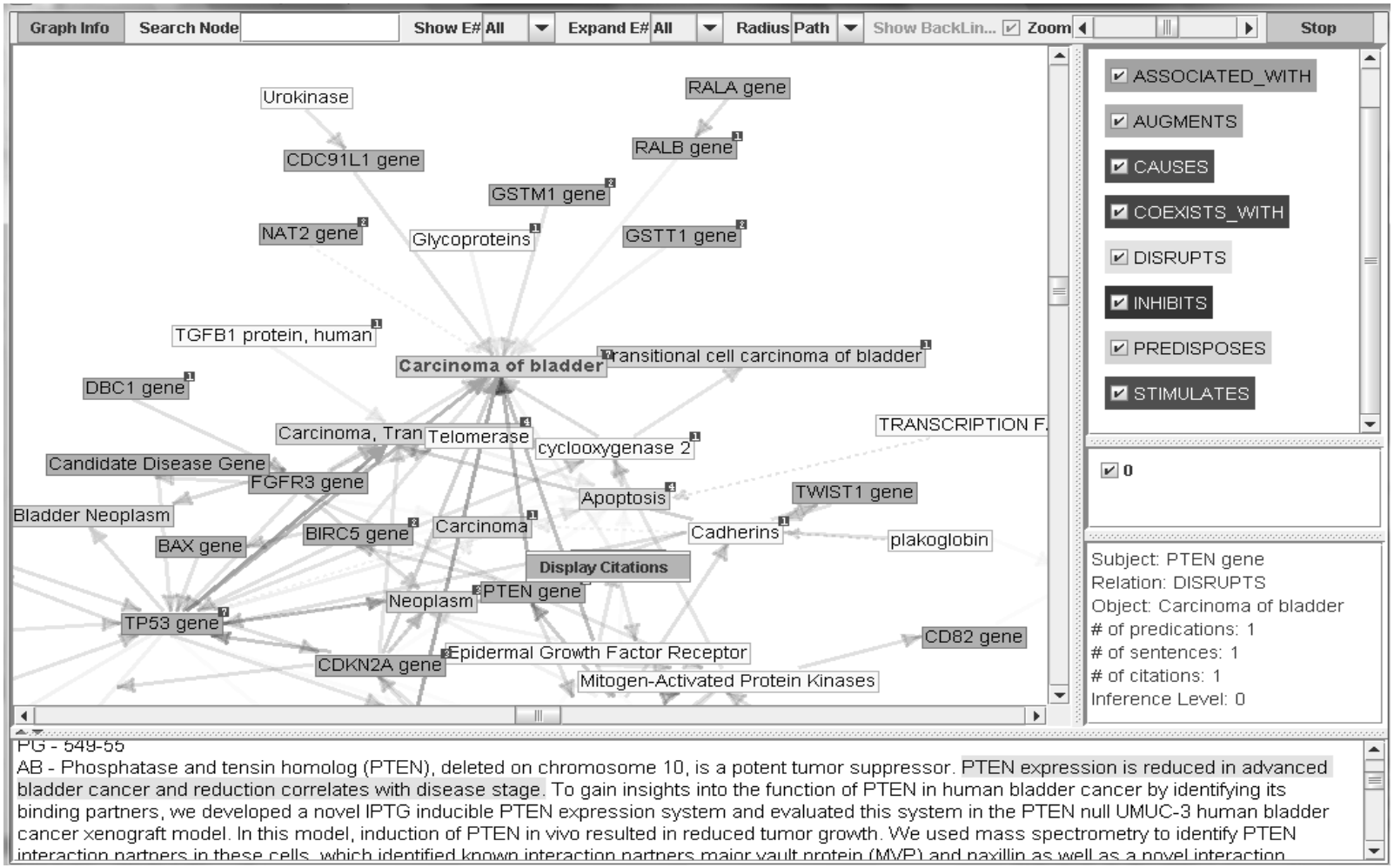

Figure 6. Visualized Summarized Results. This is an image of the Visualization process displaying summarized data addressing the genetic etiology of bladder cancer.

SemRep

SemRep, [12] an NLM rule-based symbolic natural language processing system, extracts meaning from text in citation title and abstract fields and expresses it in the form of semantic predications. For example, if the original text reads:

> "The *IGF1R* is up-regulated *in bladder cancer* compared with non-malignant bladder, and might contribute to a propensity for invasion" [13].

SemRep produces this predication:

> IGF1R gene | gngm | ASSOCIATED_WITH | Carcinoma of bladder | neop

In this example, SemRep, which integrates MetaMap [14] concept mapping functionality, has determined that "IGF1R gene" and "Carcinoma of bladder" are the respective subject and object arguments in the original text by mapping the original sentence's terms to preferred concepts in the Unified Medical Language System (UMLS) [15] Metathesaurus. These arguments are connected by a predicate, in this case "ASSOCIATED_WITH," indicating the relationship that binds them in the sentence. Additionally, SemRep identifies the semantic types within the UMLS Semantic Network associated with the arguments. In this case, IGF1R is associated with the semantic type "Gene or Genome," (abbreviated as *gngm*) and Carcinoma of bladder is associated with the semantic type "Neoplastic Process" (abbreviated as *neop*).

Summarization

Semantic MEDLINE summarization [16] filters the SemRep output, identifying the semantic predications conforming to a conceptual point-of-view, as constrained by a UMLS Metathesaurus seed topic indicated by the user. For example, a user could direct

Semantic MEDLINE to summarize for the diagnosis (point-of-view) of coronary artery disease (seed topic). Summarization filters the semantic predications from the SemRep stage in four sequential steps:

Relevance: Collects semantic predications addressing the user-selected seed topic of the summary. For example, if the user chose the UMLS Metathesaurus topic "Coronary Arteriosclerosis," summarization would collect all predications that included this seed topic as a subject or object argument.

Connectivity: Augments *Relevance* semantic predications with others which share a nonseed topic argument. In continuing the example, the schema would note that the predication "Coronary Arteriosclerosis COEXISTS_WITH Inflammation" includes the argument "Inflammation." Connectivity filtering would identify other predications which also include this argument and add them to the *Relevance* group.

Novelty: Eliminates semantic predications declaring basic knowledge which users likely know, such as "Coronary Arteriosclerosis ISA Vascular Disease(s)," by paring away such predications containing general, higher level UMLS Metathesaurus concepts.

Saliency: Limits final output to semantic predications that occur most frequently. For example, the predication "tomography DIAGNOSIS Coronary Arteriosclerosis" would be included in the final output if it occurred a sufficient number of times.

When using the Semantic MEDLINE Web-based interface, users choose the desired point-of-view and seed topic from pull-down menus. Available seed topic choices are

automatically determined by mapping UMLS Metathesaurus concepts to the SemRep data. Point-of-view choices are dependent on what *individually crafted* software applications known as schemas have been incorporated into Semantic MEDLINE. Within each schema, permitted semantic predications are restricted to a limited number of subject_predicate_object patterns, with semantic types serving as predicate arguments. For example, the schema designed for a *diagnosis* point of view permits only semantic predications containing CAUSES, DIAGNOSIS, LOCATION_OF, COEXISTS_WITH, PROCESS_OF, and ISA as predicates, and limits their arguments to a group of specifically named semantic types. Prior research in determining what predicates and semantic types best express a point-of-view enables schema designers to encode which specific semantic predication patterns the schema should seek.

Manually coded schema creation requires significant time and expertise. Research to determine relevant predicates and semantic types, plus time to code and test each schema, are required. At this time, there are only four schemas in place in the Semantic MEDLINE prototype Website enabling users to summarize according to four points of view: treatment of disease; [17] substance interaction; [18] diagnosis; [19] and pharmacogenomics [20]. A fifth schema summarizing data for a genetic etiology of disease point-of-view has been developed by one of us [TEW], [21] but has not yet been incorporated into the Internet version of Semantic MEDLINE. It is difficult to quantify how many points-of-view would be needed in order to satisfy most information needs; however, point-of-view refinement is roughly comparable to the conceptual scope of a subheading enhancement for MeSH subject headings, and currently there are 83 subheadings in use [22]. The five points-of-view substantially fall short when compared

to subheading availability for basic IR. It would take schema developers a considerable amount of time to create enough conventional schemas to provide such summarization potential in Semantic MEDLINE.

We hypothesize that the MEDLINE output based on a user-generated PubMed query that is constrained to the desired topic and point-of-view will generate SemRep output that likely contains a semantic profile representative of the same topic and point-of-view focus. Properties of SemRep output, particularly term frequencies, may indicate the topic and point-of-view expressed in the original PubMed query. Prior efforts in leveraging term and pattern frequencies have been effective in other summarization applications [23]. An algorithm leveraging SemRep output term frequencies could dynamically infer a user's information needs and summarize accordingly. This adaptive, dynamic algorithm would accommodate summarization for diverse points of view and eliminate the need for multiple schemas.

Project Aim

The aim of this project was to develop and evaluate an algorithm which utilizes statistical metrics to automatically identify predications salient to a seed topic and point-of-view as expressed in a PubMed search query. The work started out with the initial task of supporting secondary genetic database curation, but the method is general enough to apply to other tasks. The use of SemRep as a semantic predication generator is a choice of convenience. As long as there is a sufficiently representative collection of texts available for the algorithm to use in appraising data, the methods described here are

sufficiently generalizable to apply to semantic predications produced by other applications [24, 25].

Methods

We developed a new algorithm that dynamically identifies salient SemRep output, and then evaluated its utility by comparing its performance to that of a conventional summarization schema, as well as two of the individual metrics which form the algorithm. MEDLINE data was harvested via PubMed using a query expressing a specific topic and point-of-view. The citations were processed by SemRep. We summarized the SemRep output by applying the new algorithm, guided by the four-filter architecture described earlier. To assess the algorithm's collective efficiency, we also summarized the SemRep data by separately applying the algorithm's two core metrics. We also processed the SemRep output with a conventional summarization schema designed to filter data according to a genetic etiology of disease point-of-view. Bladder cancer served as the seed topic in each case. In order to evaluate outputs, we simulated the task of secondary genetic database curation. In this task, a semantic predication such as "TP53 gene | ASSOCIATED_WITH | Carcinoma of bladder" is desirable, because it offers data salient to the work of annotating gene and disease process information in a database like OMIM or GHR. The semantic predicate "Excision | TREATS | Carcinoma of bladder" is not desirable, because it offers no information addressing gene function in disease development for database curators. We extracted genetic entities (e.g., genes, proteins) from the outputs of all summarizing methods. Using a reference standard, we measured precision, recall, and F-scores for the outputs of all summarization methods.

Algorithm Development

After researching several different approaches, adaptations of the Kullback-Leibler Divergence [26] and Riloff''s RlogF metric [27, 28] demonstrated promising capabilities in identifying salient predications in SemRep output.

Kullback-Leibler Divergence

The Kullback-Leibler Divergence (KLD) determines the difference between a true distribution (P) and an assumed distribution (Q):

$$D(P||Q) = \Sigma\ P(x)\log_2(P(x)/Q(x))$$

where x represents the relative frequency of each unique predicate in each distribution. In our case, we compare the distribution of SemRep predicates resulting from a PubMed query that expresses a seed topic and a point-of-view, as distribution P, with a large dataset of predicates expressing the seed topic from all of MEDLINE (i.e., representing all points of view), as distribution Q. We compare only shared predicates. The individual KLD calculation (before summing) assigns a value to each predicate indicating its prevalence in distribution P, expressing the single point-of-view. For example, if the predicate ASSOCIATED_WITH had a relative frequency of 0.290 in distribution P and 0.076 in distribution Q, its KLD value would be 0.5603. Semantic predications in both distributions are limited to those containing a chosen UMLS Metathesaurus seed topic before KLD analysis. A database that contains SemRep output for all MEDLINE

citations published between Jan 1, 1999 to August 31, 2009 served as the data source for distribution Q.

RlogF

The RlogF metric was designed to assess the relevance of extracted patterns in unlabeled text, and was applied to SemRep output to measure the significance of semantic types as bound to a single predicate. Because the semantic type associated with the seed topic is so prevalent in the data, RlogF was adapted to assess the significance of a *nonseed topic semantic type* as bound to a predicate in each semantic predication:

$$\text{RlogF}(\text{pattern}_i) = \log_2(\text{semantic type frequency}_i) * P(\text{relevant} \mid \text{pattern}_i)$$

The conditional probability ($P(\text{relevant} \mid \text{pattern}_i)$) is the quotient of the raw frequency of a specific semantic type as bound to a given predicate, divided by the raw frequency of all semantic types as bound to the same predicate:

$$P(\text{relevant} \mid \text{pattern}_i) = \frac{\text{semantic type frequency}_i}{\text{total frequency}}$$

The pattern's conditional probability is weighted by the log of its frequency, represented here as $\log_2(\text{semantic type frequency}_i)$. For example, if the nonseed topic semantic type *gngm* occurs with the predicate ASSOCIATED_WITH 107 times, and all combined nonseed semantic types occur with the same predicate 171 times, the resulting RlogF score will be 4.22. The use of the $\log_2()$ term serves to flatten the dynamic range of the

probability space, rewarding semantic types that are very strongly correlated with the relevant patterns while still rewarding moderately correlated types that occur very frequently.

PredScal

Raw RlogF scores can exceed raw KLD scores, yet they express a different relationship in SemRep's output space. KLD scores express a proportional relationship among predicates across the entire dataset, while RlogF values express a binding between a single predicate and its associated semantic types. Therefore, we created a scaling function named PredScal to scale RlogF values according to the spatial proportions of predicates in a given dataset:

$$\text{PredScal} = 1 / \log_2(c)$$

In this metric, c represents the count of unique predicates in a dataset. For example, if there were 16 unique predicates in a dataset, PredScal would equal a scaling factor of 0.25.

The three metrics were combined to form a new algorithm, called "Combo," to evaluate SemRep data:

$$\text{Combo} = (\text{RlogF} * \text{PredScal}) * \text{KLD}$$

Each semantic predication has the form SemanticType_a $predicate_i$ $\text{SemanticType}_{b,}$ so its Combo score is calculated by scaling the RlogF metric of the predicate/nonseed topic semantic type with the PredScal metric, then multiplying the result with the predicate's KLD score.

Data

MEDLINE citations returned by PubMed for the following search query were downloaded:

("2003/01/01"[Publication Date] : "2008/07/31"[Publication Date]) AND (Urinary Bladder Neoplasms/genetics[majr] AND Urinary Bladder Neoplasms/etiology[majr]) AND English[la]

The search query focuses on the genetic etiology (the point-of-view) of bladder cancer (the topic). In this query, we limited citation output to a five-year span merely as a convenience, allowing us to utilize an existing reference standard.

Algorithm Application

We utilized Combo as the operative mechanism in the final *Saliency* filter in the four-filter architecture. To obtain results for the *Relevancy* filter, we extracted all novel semantic predications from the SemRep data which included the UMLS Metathesaurus seed topic “Carcinoma of bladder” as an argument. Then we applied the Combo algorithm to derive a score for each semantic predication. The semantic type associated with “Carcinoma of bladder” is *neop*; the nonseed semantic type associated with the

opposing subject/object argument in each semantic predication was used in performing the algorithm's RlogF calculation.

In order to explore salient predications that would result from the *Connectivity* filter, we performed a similar analysis on the novel SemRep output which did not include the UMLS Metathesaurus seed topic "Carcinoma of bladder," but did share the nonseed topic semantic type gngm with two of the highly ranked *Relevancy* predications. We also applied the Combo algorithm to derive a score for each semantic predication in this *Connectivity* group. We calculated the RlogF scores using the semantic type other than the seed gngm in deriving a Combo score for each semantic predication. In the case of these predications sharing the gngm semantic type, if their other semantic type was *neop* it was associated with UMLS Metathesaurus concepts such as "Neoplasm progression" and "Carcinoma, Transitional Cell."

To reiterate the four-filter architecture application description, we note that in both of the above procedures (i.e., *Relevance* and *Connectivity* filtering) we included only novel predications in our analyses, thus simulating *Novelty* filtering. The *Saliency* filtering phase consisted of the Combo algorithm application to identify the most informative predications.

To serve our task-based analysis, we extracted all genetic entities noted as arguments in the four top-ranked novel *Relevancy* semantic predication patterns, and the top-ranked novel *Connectivity* pattern.

Individual Metric Application

To assess the efficiency of the combined metrics in the Combo algorithm, we also separately applied the KLD and RlogF metrics in summarizing the SemRep data within the four-filter architecture. To simulate the *Relevance* stage for KLD summarization, we identified the four predicates with the highest KLD scores which included the seed topic "Carcinoma of bladder" as a subject or object argument. All novel semantic predications including these predicates and the seed topic were extracted as salient output. To simulate the *Connectivity* stage, we identified the highest scoring predicate, using the most prominent shared semantic type argument from the *Relevance* stage as the shared argument seed in KLD computation. All novel semantic predications containing the top *Connectivity* stage predicate and shared semantic type were also extracted as salient output. We extracted all genetic entities serving as subject or object arguments in the salient output.

To independently apply the RlogF metric in summarizing the SemRep output within the *Relevance* stage, we identified the four top scoring RlogF predicate / nonseed semantic type pairings among all semantic predications which included the seed topic "Carcinoma of bladder". Novel semantic predications which included these top four predicate / nonseed semantic type pairs were extracted as salient output. To simulate the *Connectivity* summarization phase, we identified the predicate / nonseed semantic type pair with the highest RlogF score among all semantic predications that contained the most prominent shared semantic type from the *Relevance* phase. Novel semantic predications containing this predicate / nonseed semantic type pair were also set aside as

salient output. We extracted all genetic entities serving as subject or object arguments in the salient output.

Conventional Schema

A conventional schema designed to summarize for the point-of-view of genetic etiology of disease also processed the SemRep data. Genetic entities serving as arguments were also extracted from its output.

Evaluation

To evaluate the four groups of extracted genetic entities, we normalized their names to coincide with the associated gene names in Entrez Gene, and compared them to a reference standard of genes implicated in bladder cancer development in selected OMIM and GHR records, originating from primary literature published between January 1, 2003 and July 31, 2008. To normalize protein, peptide, and amino acid terms, we identified the gene which exclusively produced the entity according to Entrez Gene records, and replaced each term with the matching gene name. Terms which were too general to be matched to a specific gene were discarded. The reference standard had been assembled prior to this study in order to evaluate the conventional schema [21]. One of us (TEW) and another colleague reviewed OMIM and GHR records having a major focus on bladder cancer and the genes potentially involved in its development. They identified 13 genes which had proven secondary genetic database curation appeal because of their descriptions in the OMIM and GHR records. Results for this study were evaluated in terms of recall, precision, and F-score.

Results

The base search query provided 667 citations focused on genetic etiology of bladder cancer. Leveraging MeSH indexing (i.e., the use of the *[majr]* flag in the query above) resulted in citations that included both the genetic and the etiologic factors of bladder cancer as major themes. SemRep processed the 667 citations, resulting in 5,421 semantic predications.

The four summarization methods provided diverse results in terms of raw and task-based output. The Combo summarization method identified 201 salient semantic predications, while the KLD metric alone identified 630 salient semantic predications, and the RlogF metric alone identified 177 salient semantic predications. The conventional schema identified 112 salient semantic predications. The top-ranking novel *Relevance* and *Connectivity* predication scores from the Combo, KLD, and RlogF analyses are listed in Tables 4 - 6. There were 74 individual genes identified as implicated in bladder cancer development in the Combo output. The KLD metric alone identified 77 genes, and the RlogF metric alone identified 69 genes implicated in bladder cancer development. The conventional schema output included 10 such implicated genes.

Recall for the four summarization methods was calculated by comparing outputs to the reference standard of genes noted in relevant GHR and OMIM records as noteworthy in bladder cancer development. Summarization using the Combo algorithm achieved 61% recall. The KLD and Rlogf summarization methods also achieved 61% recall. The conventional schema achieved 23% recall. The reference standard includes genes implicated in bladder cancer development in specific GHR and OMIM records, but

Table 4. . Combo Scores of Top-Ranking Patterns in Novel Relevance and Novel Connectivity Analyses; nonseed semantic types are indicated in square brackets.

Relevancy Analysis Seed Topic: Carcinoma of bladder	**Combo Score**
[gngm] ASSOCIATED_WITH neop	0.592531
[gngm] PREDISPOSES neop	0.205778
[aapp] ASSOCIATED_WITH neop	0.152883
[aapp] PREDISPOSES neop	0.039868
Connectivity Analysis Shared Semantic Type: gngm	**Combo Score**
gngm ASSOCIATED_WITH [neop]	0.873016

Table 5. Kullback-Leibler Divergence Scores of Top-Ranking Predicates in Novel Relevance and Novel Connectivity Analysis.

Relevance Analysis Seed Topic: Carcinoma of bladder	**KLD Score**
ASSOCIATED_WITH	0.561861059
PREDISPOSES	0.299181776
AFFECTS	0.088951936
PART_OF	0.034851914
Connectivity Analysis Shared Semantic Type: gngm	
ASSOCIATED_WITH	0.5553145

Table 6. RlogF Scores of Top-Ranking Predicate / Nonseed Semantic Type Pairs in Novel Relevance and Novel Connectivity Analysis

Relevance Analysis Seed Topic: Carcinoma of bladder	**RlogF Score**
gngm ASSOCIATED_WITH	4.218344839
topp TREATS	2.96127605
ISA neop	2.807354922
gngm PREDISPOSES	2.751207824
Connectivity Analysis Shared Semantic Type: gngm	
ASSOCIATED_WITH neop	7.208071323

likely does not represent a comprehensive list of genes associated with bladder cancer development. The reference standard provides a list of genes whose value has already been confirmed within the task of secondary genetic database curation, because GHR and OMIM curators have annotated their potential roles in bladder cancer development. The results of the reference standard analysis are listed in Table 7.

Precision was evaluated by taking the previously established true positive findings into account with the additional genes included as arguments in the semantic predications identified as salient by the four summarization methods. To assess validity (true positive or false positive status) for the additional genes, Genes into Reference (GeneRIF) notations in relevant Entrez Gene records were reviewed for disease process implication, thus confirming appeal for the simulated task of genetic database curation. If the relevant Entrez Gene record did not contain applicable GeneRIFs, but otherwise noted bladder cancer association, the gene was assigned true positive status. Summarization with the new Combo algorithm achieved 81% precision. The KLD summarization method attained 70% precision, and the RlogF method achieved 72% precision. The conventional schema attained 100% precision. Table 8 highlights precision scores.

We calculated F-scores for each method to assess a balance between recall and precision. The Combo summarization method resulted in an F-score of 0.69. The KLD and RlogF methods yielded F-scores of 0.65 and 0.66, respectively. Summarization with the conventional schema produced an F-score of 0.37.

Table 7. Recall Results with Reference Standard (TP=True Positive; FN=False Negative)

Gene	Combo Analysis	KLD Analysis	RlogF Analysis	Conventional Schema
FGFR3	TP	TP	TP	TP
XPD	TP	TP	TP	FN
RAG1	FN	FN	FN	FN
TP53	TP	TP	TP	TP
MTCYB	FN	FN	FN	FN
HRAS	TP	TP	TP	FN
NAT2	TP	TP	TP	TP
RB1	TP	TP	TP	FN
TSC1	TP	TP	TP	FN
ATM	FN	FN	FN	FN
TGFB1	FN	FN	FN	FN
MDM2	TP	TP	TP	FN
ERBB3	FN	FN	FN	FN
Recall	61%	61%	61%	23%

Table 8. Precision Results (TP=True Positive; FP=False Positive)

	Combo Analysis	KLD Analysis	RlogF Analysis	Conventional Schema
TP	60	54	50	10
FP	14	23	19	0
Total	74	77	69	10
Precision	81%	70%	72%	100%

Discussion

In this study's task-based context (i.e., genetic database curation), summarization with the new Combo algorithm outperformed the conventional schema in terms of raw output and recall, while maintaining reasonable precision. Combo also produced a higher F-score than the separate KLD and RlogF applications, thus attaining a slightly superior balance of recall and precision. All of the five patterns that Combo identified as salient (Table 4) yielded semantic predications containing gene arguments, with an average of 26 separate arguments per pattern. In the separate KLD application, the predicates AFFECTS and PART OF, when paired with the seed topic in the *Relevance* phase, together produced only nine gene arguments while all salient KLD patterns (Table 5) produced an average of 32 separate arguments. The nine arguments produced by AFFECTS and PART_OF were duplicated elsewhere in the KLD analysis. Each RlogF pattern (Table 6) produced an average of 22 separate gene arguments. Semantic predications matching the two RlogF salient patterns *Therapeutic or Preventive Procedure TREATS* (topp TREATS) and *ISA Neoplasm* (ISA neop) in *Relevance* summarization did not have gene arguments, and were therefore unproductive. The Combo, KLD, and RlogF applications performed identically in terms of recall; each method produced the same genes from the reference standard. Combo outperformed the separate KLD and RlogF applications in terms of precision. It produced more genes with validated curation potential.

Because the Combo algorithm is designed to adaptively identify relevant data through analysis of a SemRep dataset's individual properties, its generalizability gives it potential to address information needs other than genetic disease etiology. The algorithm could be

encoded into a very flexible schema for integration into the Semantic MEDLINE model. The new dynamic schema could potentially enable Semantic MEDLINE to summarize for many points of view, thus transforming it into a dynamic NLP application for a diverse range of needs. The Visualization component in Semantic MEDLINE would provide a graphical representation of the summarized results (see an example of visualized genetic etiology of bladder cancer findings in Figure 6).

There are several information needs that a dynamic schema could address. Secondary database curators could implement it in order to find additional genes associated with a disease process, as recorded in bibliographic text. Researchers in other fields may also benefit from dynamic text summarization. The following vignettes illustrate Combo's generalizability by exploring how Semantic MEDLINE, empowered by this new algorithm, may benefit multiple information needs.

Primary Research

In the initial work of research, scientists usually review prior studies related to a planned investigation. This can be a time-consuming step. For example, scientists exploring the causes of myocardial infarction in humans must review over 17,000 major studies found in PubMed. Semantic MEDLINE with the Combo algorithm could facilitate this type of data appraisal. For example, researchers could execute the following query:

myocardial infarction/etiology[Majr] Limits: Humans

Then, they could choose the UMLS Metathesaurus seed topic(s) addressing their needs. Results would then be reviewed using the graphic display, giving an immediate overview

of salient content. The researchers could execute searches limited by time ranges (e.g., items published within a three-year period) to simplify the amount of data within the Visualization graph, and to note how research chronologically evolved in the field. Effective use of Semantic MEDLINE as a research appraisal tool could accelerate investigational studies and eventually quicken the bench to bedside process in clinical care.

Clinical Decision Support

Online biomedical databases such as MEDLINE can answer clinicians' questions, but are time-consuming to use [29]. Semantic MEDLINE with the Combo algorithm could quickly summarize large amounts of citations and provide a graphic representation of data addressing many information needs. Consider the following scenario: a physician assistant (P.A.) wants to prevent future injury to an elderly patient experiencing recurrent hip fractures. The P.A. submits the search "Hip Fractures[mesh] AND recurrent" and then chooses "Hip fractures" as the UMLS Metathesaurus seed topic. Using the graphic display, the P.A. notes that dementia [30] is associated with recurrent hip fracture. The P.A. realizes that addressing this comorbidity may prevent future fractures. Sorting through the citations by hand would have required too much time to be practical; acquiring the information using Semantic MEDLINE takes less than a minute.

Systematic Reviews

Evidence-Based Medicine (EBM) is "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [31].

Research based on EBM principles provides scientifically grounded information for patient care. Consider the following scenario: a working group within the Cochrane Collaboration wishes to update a review offering dietary advice for cardiovascular disease reduction [32]. They compose and execute the following PubMed query:

diet[mesh] AND cardiovascular diseases/prevention and control[mesh]

The search is limited to Randomized Controlled Trials, which results in 432 citations.

The group uses Semantic MEDLINE with the Combo algorithm to assess the citations, choosing the most relevant seed topics. This provides them with an immediate visual assessment of the randomized controlled trials, giving them a starting point in evaluating and selecting research to include in their systematic review.

Limitations

This study compared conventional schema output to the statistical algorithm's performance in the context of the single task of secondary genetic database curation for the genetic etiology of bladder cancer. We cannot quantify its performance in other applications until similar research determines it. However, Semantic MEDLINE with conventional summarization has proven to be effective in identifying evidence-based treatment of 50 diseases [17]. Considering the overall performance improvement demonstrated by the new statistical algorithm over traditional summarization, it also holds promise in other applications. In conducting this study, we did not have access to curators' individual protocol and thought processes, which are clearly essential to know in a real-world curation application of Combo. We can, however, speculate on what information is valuable in database curation by what is noted in the biomedical literature.

We should also note that Summarization performance in the Semantic MEDLINE model is dependent on the query results, specifically, the search query's performance, the quality of the citations garnered in IR, and SemRep's accuracy in capturing the citations' content.

Conclusion

In this paper we described the development of a statistically based algorithm known as Combo that automatically summarizes SemRep semantic predications for a topic and a point-of-view in the Semantic MEDLINE model. We evaluated summarization utilizing Combo by comparing it to conventional summarization, using a previously established reference standard, in the task-based context of secondary genetic database curation. We also proposed real-world scenarios showing how Semantic MEDLINE, empowered with the new Combo algorithm, could benefit additional information needs. Combo is not limited to predications generated by SemRep; any predication generator that produces subject_predicate_object triplets could benefit from Combo.

Abbreviations

IR: Information Retrieval; NLM: National Library of Medicine; NLP: Natural Language Processing; GHR: Genetics Home Reference; OMIM: Online Mendelian Inheritance in Man; P.A.: Physician Assistant; UMLS: Unified Medical Language System

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

TEW developed the Combo algorithm, designed the study, harvested the data, oversaw data processing with SemRep and the conventional genetics schema, performed the Combo, KLD, and RlogF summarizations, conducted the evaluation, and wrote the original manuscript. JFH provided essential manuscript revisions, project supervision, and mentoring, and suggested the use of the RlogF metric. Both authors read and approved the final manuscript.

Acknowledgements and Funding

We wish to acknowledge Thomas Rindflesch and Marcelo Fiszman for their essential work in developing SemRep and the Summarization concept. We also wish to thank the National Library of Medicine for its support and for funding this work through grant number T15LM007123. JFH's effort in this work was supported in part by 5R21LM009967-02. The NIH played no role in the collection, analysis, interpretation of data; the writing of the manuscript; or the decision to submit the manuscript for publication.

References

1. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998 Oct 21;280(15):1347-52.

2. Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J. 2006 Mar;23(1):3-12.

3. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao ZX, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang LL, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003 Oct;13(10):2363-71.

4. Fraser C, Murray A, Burr J. Identifying observational studies of surgical interventions in MEDLINE and EMBASE. BMC Med Res Methodol. 2006;6:41.

5. Bekhuis T, Demner-Fushman D. Towards automating the initial screening phase of a systematic review. Stud Health Technol Inform. 2010;160(Pt 1):146-50.

6. Karimi S, Pohl S, Scholer F, Cavedon L, Zobel J. Boolean versus ranked querying for biomedical systematic reviews. BMC Med Inform Decis Mak. 2010;10.

7. Haase A, Follmann M, Skipka G, Kirchner H. Developing search strategies for clinical practice guidelines in SUMSearch and Google Scholar and assessing their retrieval performance. BMC Med Res Methodol. 2007;7:28.

8. Lin Y, Li W, Chen K, Liu Y. A document clustering and ranking system for exploring MEDLINE citations. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):651-61.

9. Niu Y, Zhu X, Hirst G. Using outcome polarity in sentence extraction for medical question-answering. AMIA Annu Symp Proc. 2006:599-603.

10. Johnson DB, Zou Q, Dionisio JD, Liu VZ, Chu WW. Modeling medical content for automated summarization. Ann N Y Acad Sci. 2002 Dec;980:247-58.

11. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC. Semantic MEDLINE: a web application for managing the results of PubMed searches Proceedings of the Third International Symposium for Semantic Mining in Biomedicine, 2008: 69-76.

12. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller S, Hersh W, Friedman C, eds. Medical informatics: knowledge management and data mining in biomedicine: Springer, 2005: 399-422.

13. Rochester MA, Patel N, Turney BW, Davies DR, Roberts IS, Crew J, Protheroe A, Macaulay VM. The type 1 insulin-like growth factor receptor is over-expressed in bladder cancer. BJU Int. 2007 Dec;100(6):1396-401.

14. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

15. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

16. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, 2004:76-83.

17. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.

18. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc. 2006:254-8.

19. Sneiderman C, Demner-Fushman D, Fiszman M, Rosemblat G, Lang FM, Norwood D, Rindflesch TC. Semantic processing to enhance retrieval of diagnosis citations from Medline. AMIA Annu Symp Proc. 2006:1104.

20. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

21. Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc. 2010;98(4):273 - 81.

22. Qualifiers - 2011. Bethesda: National Library of Medicine, 2010. [rev. 30 August 2010 2010; cited 1 October 2010 2010]. <http://www.nlm.nih.gov/mesh/topsubscope.html>.

23. Hahn U, Mani I. The challenges of automatic summarization. Computer 2000;33(11):29 - 36.

24. Khoury R, Karray F, Sun Y, Kamel M, Basir O. Semantic understanding of general linguistic items by means of fuzzy set theory. IEEE Transactions on Fuzzy Systems 2007;15(5):757 - 71.

25. Cole S, Royal M, Valtorta M, Huhns M, Bowles J. A lightweight tool for automatically extracting causal relationships from text. Proceedings of the IEEE SoutheastCon 2006. Memphis, TN: IEEE, 2005.

26. Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics 1951;22(1):79 – 86.

27. Riloff E. Automatically generating extraction patterns from untagged text. Proceedings of the Thirteenth National Conference on Artificial Intelligence. Menlo Park, CA: The AAAI Press/MIT Press, 1996: 1044–9.

28. Riloff E, Phillips W. An introduction to the Sundance and Autoslog Systems. University of Utah School of Computing, 2004. Report No.: UUCS-04-015.

29. Chambliss ML, Conley J. Answering clinical questions. J Fam Pract. 1996 Aug;43(2):140-4.

30. Yamanashi A, Yamazaki K, Kanamori M, Mochizuki K, Okamoto S, Koide Y, Kin K, Nagano A. Assessment of risk factors for second hip fractures in Japanese elderly. Osteoporos Int. 2005 Oct;16(10):1239-46.

31. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996 Jan 13;312(7023):71-2.

32. Brunner EJ, Rees K, Ward K, Burke M, Thorogood M. Dietary advice for reducing cardiovascular risk. Cochrane Database Syst Rev. 2007(4):CD002128.

CHAPTER 4

RETHINKING INFORMATION DELIVERY: USING A NATURAL LANGUAGE PROCESSING APPLICATION FOR POINT-OF-CARE DATA DISCOVERY

T. Elizabeth Workman, M.L.I.S., Ph.D.c.

Joan M. Stoddart, M.A.L.S.

Journal of the Medical Library Association: JMLA (in Press)

Abstract

Objective: This paper examines the use of Semantic MEDLINE, a natural language processing application enhanced with a statistical algorithm known as Combo, as a potential decision support tool for clinicians. Semantic MEDLINE summarizes text in PubMed citations, transforming it into compact declarations that are filtered according to a user's information need that can be displayed in a graphic interface. Integration of the Combo algorithm enables Semantic MEDLINE to delivery information salient to many diverse needs. **Methods:** The authors selected three disease topics, and crafted PubMed search queries to retrieve citations addressing the prevention of these diseases; they then processed the citations with Semantic MEDLINE, with the Combo algorithm enhancement. To evaluate the results, they constructed a reference standard for each disease topic consisting of preventive interventions recommended by a commercial decision support tool. **Results:** Semantic MEDLINE with Combo produced an average recall of 79% in primary and secondary analyses, an average precision of 45%, and a final average f-score of 0.57. **Conclusion:** This new approach to point-of-care information delivery holds promise as a decision support tool for clinicians. Health sciences libraries could implement such technologies to deliver tailored information to their users.

Introduction

Clinicians often encounter information needs in their work of caring for patients. In their 2005 study, Ely and his colleagues discovered that physicians developed an average of 5.5 questions for each half-day observation, yet could not find answers to 41% of the

questions for which they pursued answers [1]. Ely cites time constraints as one of the barriers preventing clinicians from finding answers. In another study, Chambliss and Conley also found that answer discovery is excessively time consuming [2].

Chambliss and Conley determined that MEDLINE data could fulfill or nearly fulfill 71% of clinicians' answerable questions; however, PubMed is an impractical tool for point-of-care information delivery. It generally returns excessive, irrelevant data, even when implementing diverse search strategies [3]. Clinicians can spend an average of 30 minutes answering a question using MEDLINE data [4]. This is by and large due to the process of literature appraisal, which is naturally lengthened by excessive retrieval [5]. This information discovery process is not practical for a busy clinical setting [4].

Semantic MEDLINE

Natural language processing (NLP) applications such as Semantic MEDLINE can filter PubMed text for a user's specific need and summarize it to facilitate literature appraisal [6]. Semantic MEDLINE, a resource developed by the National Library of Medicine (NLM), if enhanced by an adaptive algorithm known as Combo [7], can simplify MEDLINE data for many information needs. The user activates the Semantic MEDLINE application by submitting a search query expressing his or her information need to PubMed. Semantic MEDLINE then uses the individual processes of SemRep, Summarization, and Visualization to quickly transform the citations' title and abstract text into a compact form and identify data which is salient to a specific information need, which then can be displayed in a visual graph. The following text describes these individual processes. Currently, NLM hosts the only Semantic MEDLINE application.

This study evaluated an enhanced Semantic MEDLINE system that accommodates additional information needs; this paper also briefly describes how an organization could develop it to serve its own users.

SemRep

SemRep [8], a rule-based NLP application within Semantic MEDLINE, interprets the meaning of PubMed title and abstract text, and rephrases it into compact declarations called semantic predications. For example, consider the following citation title text:

> "**Taurolidine** is effective in the **treatment** of central venous catheter-related bloodstream **infections** in cancer patients [9]."

SemRep rephrases the text with this semantic predication:

> Taurolidine_TREATS_infection

SemRep identifies "taurolidine" and "infections" as the respective subject and object of the text, and maps them to the UMLS [10] Metathesaurus preferred concepts *Taurolidine* and *infection*. It also recognizes "treatment" as the concept that binds the subject and object terms, mapping it to the predicate TREATS, as found in the UMLS Semantic Network. SemRep also identifies the logical UMLS semantic group classifications associated with the arguments, which in this case are "Pharmacologic Substance" (associated with *Taurolidine*) and "Disease or Syndrome" (associated with *infection*).

Summarization

Semantic MEDLINE's Summarization phase identifies SemRep semantic predications that are relevant to a user's indicated information need. This process begins

by prompting the user to select a topic from a list of UMLS Metathesaurus preferred concepts that appear in the SemRep data. A summarization software application within Semantic MEDLINE processes the SemRep output according to the following sequential phases:

Relevance: Gathers semantic predications containing the user-selected seed topic. For example, if the chosen topic were Septicemia, this filter would collect the semantic predication Blood culture_DIAGNOSES_Septicemia.

Connectivity: Augments *Relevance* predications with those which share a nonseed argument's semantic type. For example, in the above predication Blood culture_DIAGNOSES_Septicemia, the semantic type of the nonseed argument Blood culture is "Laboratory Procedure". This filter would augment the *Relevance* semantic predications with others such as Measurement of serum lipid level _DIAGNOSES_Sepsis of the newborn, because "Laboratory Procedure" is also the semantic type of the subject argument Measurement of serum lipid level.

Novelty: Eliminates vague predications, such as pharmaceutical preparation_TREATS_patients, that present information that users already likely know, and are of limited use.

Saliency: Limits final output to predications that occur with adequate frequency. For example, if Blood culture_DIAGNOSES_Septicemia occurred enough times, all occurrences would be included in the final output.

To operationalize the final *Saliency* phase, the summarization software in this study used a statistical algorithm known as Combo. Combo [7] analyzes predicate frequencies using an adaptation of the Kullback-Leibler Divergence, [11] and measures the strength of

predicate/semantic type pairings with Riloff's RlogF metric [12] and PredScal, a scaling metric developed for the Combo algorithm. Prior to this approach, summarization was dependent on conventional, static applications called schemas limited to specified subject_predicate_object patterns. A different schema was required to summarize for each subheading-type refinement, limiting use to five options: treatment of disease [13], substance interaction [14], diagnosis [15], pharmacogenomics [16], and genetic etiology of disease [17]. Because of its advanced computational methodology, Combo adapts to the properties of each set of SemRep output in determining what is relevant to the user's information need, thus enabling summarization for many subheading concepts.

Visualization

The semantic predications produced by the Summarization phase can be visually displayed. Figure 7 presents an interface used by NLM to display Summarization output. Due to the nature of the data's compact structure, users can quickly focus on desired data. For example, in Figure 7 the Summarization seed topic is Septicemia, and the user has limited displayed output to items containing the predicate DIAGNOSES. In Figure 8 the user has clicked on the arc connecting Septicemia and blood culture, and is presented with the citations addressing blood culture's use as a diagnostic tool for septicemia.

Objective

The objective of this study was to evaluate the effectiveness of Semantic MEDLINE, with the statistical Combo algorithm enhancement, in identifying decision support

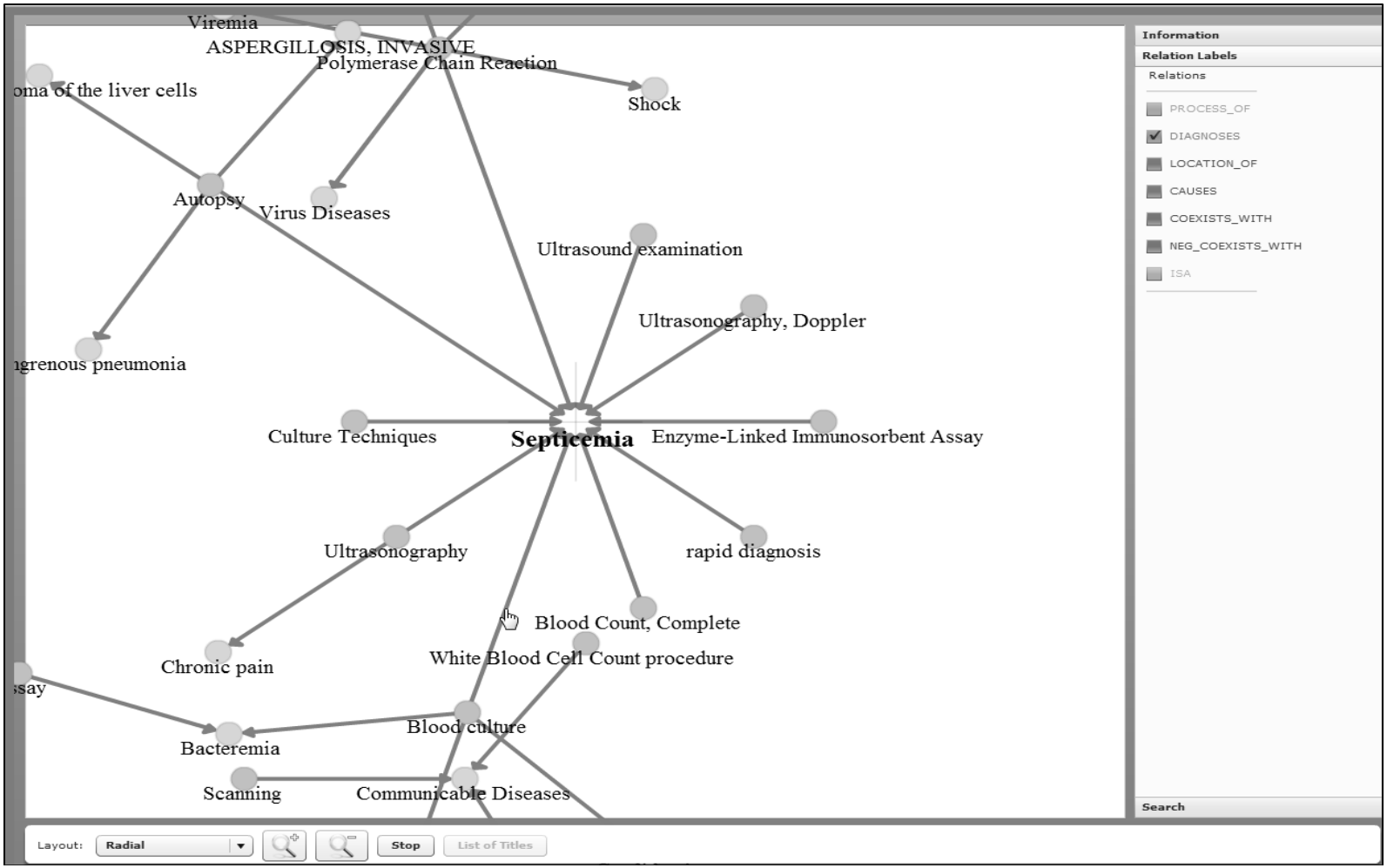

Figure 7. Summarization output

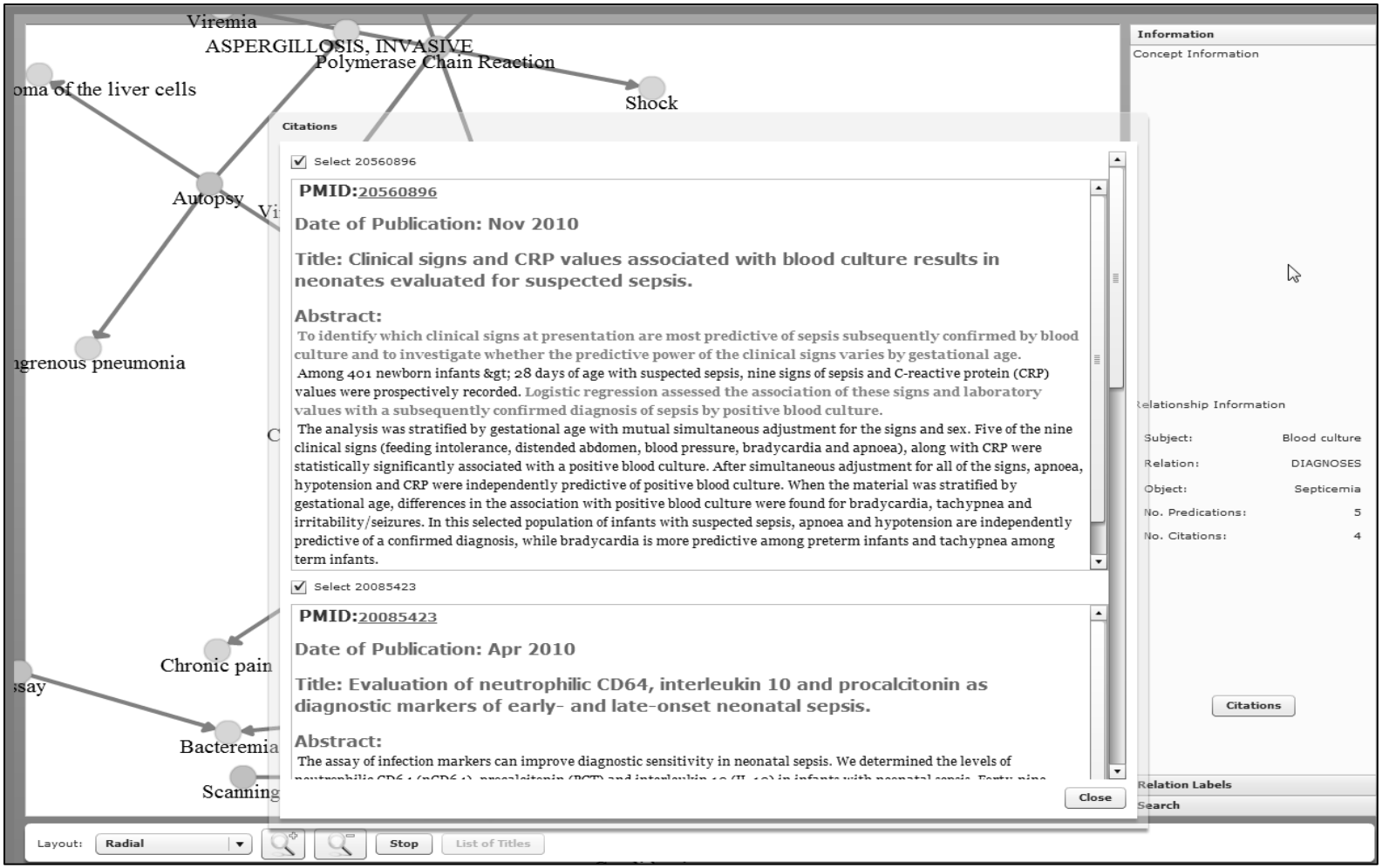

Figure 8. Summarization output with citation data

information for disease prevention. The authors wanted to explore its potential use as a point-of-care information delivery application. They wanted to determine if this approach could retrieve recommended preventive interventions found in a commercial, manually-annotated product. Prior efforts in applying Semantic MEDLINE, with the Combo algorithm, to identify information relevant to genetic disease etiology were successful, within a simulated database curation task [7]. The authors wanted to evaluate the system within a simulated clinical decision support task.

The authors wanted to evaluate this system's performance in retrieving prevention information, because the concept is fluid, and especially difficult to capture with such an NLP approach. For example, preventing congestive heart failure includes treating hypertension in vulnerable patients. To prevent lung cancer, clinicians counsel patients on smoking cessation. Therefore, the authors hypothesized that, in addition to finding relevant output in the form of "Intervention X _PREVENTS_Disease Y", they would also find relevant semantic predications containing other predicates, such as TREATS. Currently, there is no conventional static schema in NLM's Semantic MEDLINE designed to accommodate a disease prevention subheading refinement. The results of this study may offer commentary on the potential enhancement offered by Combo-driven Summarization in expanding Semantic MEDLINE's functionality.

This study also served as a pilot for a larger project to examine Semantic MEDLINE's efficiency, when enhanced with the Combo algorithm, in aiding decision support for disease prevention and drug treatment.

Methods

Disease Topics and Data

The authors chose the three topic diseases *acute pancreatitis*, *coronary artery disease*, and *malaria*. These three diseases have various etiologies, and call for a variety of types of preventive interventions. These differences in disease characteristics motivated their selection. The authors executed the following PubMed searches and downloaded the resulting citations:

Acute Pancreatitis Search Session:

#11 Search **#8 OR #9**
#9 Search **(pancreatitis/prevention and control[mesh] NOT Pancreatitis, Chronic[mesh]) AND "systematic review"** Limits: **Review, Publication Date to 2010/08/31**
#8 Search **pancreatitis/prevention and control[mesh] NOT Pancreatitis, Chronic[mesh]** Limits: **Clinical Trial, Meta-Analysis, Randomized Controlled Trial, Publication Date to 2010/08/31**

Coronary Artery Disease Search Session:

#13 Search **#10 OR #11**
#11 Search **coronary artery disease/prevention and control[mesh] AND "systematic review"** Limits: **Review, Publication Date to 2010/10/31**
#10 Search **coronary artery disease/prevention and control[mesh]** Limits: **Clinical Trial, Meta-Analysis, Randomized Controlled Trial, Publication Date to 2010/10/31**

Malaria Search Session:

#15 Search **#12 OR #13**
#13 Search **Malaria/prevention and control[mesh] AND "systematic review"** Limits: **Review, Publication Date to 2010/10/31**
#12 Search **Malaria/prevention and control[mesh]** Limits: **Clinical Trial, Meta-Analysis, Randomized Controlled Trial, Publication Date to 2010/10/31**

The search sessions were conducted February 7th, 2011. To garner evidence-based data, retrieval was focused on clinical trials, meta-analyses, randomized controlled trials,

and systematic reviews. Retrieval was also limited to match the time period represented by the study's evaluative reference standards, as described below. There were two rationales behind the search queries' structure. In evaluating Combo-enhanced Semantic MEDLINE for other related projects (addressing genetic disease etiology and drug treatment) information retrieval for text summarization was based on a single disease topic, paired with a subheading-type concept, while drawing on all citations within the database (instead of selected intricate subsets). This provided some standardization across all projects. Researchers accomplished this by combining MeSH terms with subheadings, and keyword phrases (e.g., "systematic reviews") and publication types when needed. Additionally, this specific study simulated a task in which a clinician would create the search query. Realistically, clinicians' searching skills vary, and one could expect to see anything from a very general keyword search to a more sophisticated search profiting from many of the PubMed value-added search tools. The search queries employed represented a type of middle ground in this spectrum.

Semantic MEDLINE Processing

The citations were processed with SemRep; SemRep output was processed with the Combo algorithm-enhanced Summarization application. The authors selected the following UMLS Metathesaurus preferred concepts as seed topics for the Summarization phase:

- Pancreatitis (for the acute pancreatitis citations)
- Coronary Arteriosclerosis and Coronary heart disease (for the coronary artery disease citations)

- Malaria (for the malaria citations)

Evaluation

To evaluate the results, the authors compiled a reference standard for each disease, consisting of preventive interventions recommended by DynaMed, a commercial decision support product. The authors chose DynaMed because it was one of three top-ranked products in a recent study [18], presented information in a straight-forward bullet structure, and was readily available. Preventive interventions prefaced with text such as "controversial or not well established with evidence" were not included in the study's reference standards. As previously mentioned, the authors noted the most recently published primary articles DynaMed used in identifying recommendations and limited citation retrieval in order to avoid including data published after DynaMed's source references. This approach to data acquisition was used in a similar study conducted by other investigators [13]. One of the authors (TEW) captured DynaMed data addressing prevention of the three disease topics February 6th, 2011.

The primary analysis examined Semantic MEDLINE output in the general form "Intervention X_PREVENTS_Disease Y" for Summarized output for each of the three disease topics groups, along with the associated citation from which each semantic predication originated. If a citation's text confirmed the retrieval of a reference standard intervention, it was counted as a true positive. For example, if the citation included wording such as "[the intervention] is recommended for prevention of [the disease]", the intervention received a true positive status. Knowing the nature of UMLS metathesaurus preferred concepts, the authors determined that if a general term was associated with

citation text containing a reference standard intervention's precise wording, the reference standard would receive a true positive status (this is also demonstrated in the RESULTS section). The authors limited the primary analysis to examining output in the form of "Intervention X_PREVENTS_Disease Y" because if a clinician were to use Semantic MEDLINE as a decision support tool for preventive care, he or she would likely begin by reviewing data with the PREVENTS predicate. Findings were measured according to recall, precision, and F-score. Precision scores were calculated in the primary analysis by grouping the interventions in the summarized data by name, and assessing what percentage of these groups led to related citation text containing a reference standard intervention.

The secondary analysis examined semantic predications which included predicates other than PREVENTS. The authors used the same strategy of using the associated citation data to confirm a given reference standard intervention's true positive status. Since the authors' primary interest was whether this additional data supplied additional reference standard interventions, these findings were factored into the final recall calculation.

Results

Data Acquisition and Processing

One of the authors (TEW) performed the information retrieval phase, SemRep processing, and Summarization processing using the Combo algorithm-enhanced software. The three PubMed search sessions retrieved a total of 3276 citations; the acute pancreatitis session produced 156 citations, while the coronary artery disease and malaria

sessions respectively yielded 2440 and 680 citations. SemRep produced 999 semantic predications using the acute pancreatitis citations, 14781 semantic predications from the coronary artery disease citations, and 3374 semantic predications from the 680 malaria citations. Using the associated SemRep disease topic outputs, Summarization identified 1397 unique semantic predications salient to the "Coronary Arteriosclerosis" and "Coronary heart disease" seed topics, 178 semantic predications salient to the "Pancreatitis" seed topic, and 389 semantic predications salient to the "Malaria" seed topic.

Evaluation - Primary Analysis

Semantic MEDLINE with the Combo algorithm enhancement produced an average recall of 70% in the initial examination of output in the form of "Intervention X_PREVENTS_Disease Y". The average precision was 45%, resulting in an F-score of 0.54. The primary analysis recall results for each disease topic are listed in Tables 9 - 11. Precision results are indicated in Table 12.

Evaluation - Secondary Analysis

Examination of output semantic predications containing predicates other than PREVENTS identified additional reference standard interventions, and increased average recall to 79%, with an adjusted F-score of 0.57. Reference standard results for each disease topic group are listed in Tables 9 - 11. Because all reference standard interventions for acute pancreatitis appeared in the primary analysis, no secondary analysis was necessary for this disease topic.

Table 9. DynaMed Preventive Intervention Reference Standard Recall Results, Acute Pancreatitis (TP = True Positive; FN = False Negative; N/A = Not Applicable, Found in Primary Analysis)

Intervention	**Primary Analysis**	**Secondary Analysis**
guidewire cannulation	TP	N/A
nonsteroidal anti-inflammatory drugs (NSAIDs)	TP	N/A
octreotide	TP	N/A
prophylactic nitroglycerin	TP	N/A
interleukin 10 (IL-10)	TP	N/A
	Recall: 100%	

Table 10. DynaMed Preventive Intervention Reference Standard Recall Results, Coronary Artery Disease (TP = True Positive; FN = False Negative; N/A = Not Applicable, Found in Primary Analysis)

Intervention	**Primary Analysis**	**Secondary Analysis**
proper diet	TP	N/A
aerobic exercise	FN	FN
smoking cessation	FN	TP
modifiable lifestyles	TP	N/A
weight loss	TP	N/A
treatment of diabetes	FN	TP
treatment of Hypertension	TP	N/A
treatment of Hyperlipidemia	TP	N/A
prophylactic low-dose aspirin	TP	N/A
use of ACE inhibitors	TP	N/A
complete avoidance of tobacco smoke	FN	FN
angiotensin receptor blockers	TP	N/A
aldosterone blockade	FN	FN
beta blockers	TP	N/A
influenza vaccine	FN	FN
	Recall: 60%	**Recall: 73%**

Table 11. DynaMed Preventive Intervention Reference Standard Recall Results, Malaria (TP = True Positive; FN = False Negative; N/A = Not Applicable, Found in Primary Analysis)

Intervention	**Primary Analysis**	**Secondary Analysis**
long-sleeves	FN	FN
long pants	FN	FN
window screens	FN	FN
mosquito nets	TP	N/A
insecticed-treated clothes	FN	FN
insecticed-treated nets	TP	N/A
insect repellent	TP	N/A
indoor spraying	FN	FN
insecticide treatment of livestock	FN	FN
atovaquone/proguanil	TP	N/A
trimethoprim-sulfamethoxazole	FN	FN
“antimalarial agents”	TP	N/A
artesunate plus amodiaquine or sulfadoxine-pyrimethamine	FN	TP
mefloquine	TP	N/A
sulfadoxine-pyrimethamine	TP	N/A
amodiaquine	TP	N/A
pyrimethamine plus dapsone	FN	TP
routine malaria chemoprophylaxis (i.e. during pregnancy)	TP	N/A
chloroquine	TP	N/A
recombinant vaccine based on fusion of circumsporozoite protein and HBsAg	FN	FN
RTS,S/AS02 (vaccine)	FN	FN
RTS,S/ASO2A (vaccine)	TP	N/A
RTS,S/AS01E (vaccine)	FN	TP
RTS,S/AS02D (vaccine)	FN	TP
MSP/RESA (vaccine)	TP	N/A
vitamin A supplementation	TP	N/A
	Recall: 50%	**Recall: 65%**

Table 12. Precision Results by Disease Topic, from Primary Analysis of Data Using DynaMed Reference Standards

Disease Topic	**Precision**
Acute Pancreatitis	29%
Coronary Artery Disease	45%
Malaria	61%
Average Precision	**45%**

Discussion

Findings of Two Analyses

Interesting patterns emerged from both analyses. In the primary analysis (examining output in the form "Intervention X_PREVENTS_disease Y"), of the 27 true positive findings for all three disease topics, 18 were pharmaceutical-type substances or supplements within the associated reference standards. The additional nine true positives consisted of other types of interventions, ranging from behavior issues (e.g., diet) to therapeutic technique (e.g., guidewire cannulation). In this study, Semantic MEDLINE with the Combo algorithm enhancement was more efficient at expressing preventive drug and supplement interventions with the PREVENTS predicate than for other kinds of interventions.

The secondary analysis confirmed the hypothesis that some reference standard interventions would be expressed with predicates other than PREVENTS. The secondary analysis found two of the six interventions not found in the primary analysis for coronary artery disease, and four of the 13 interventions not located for malaria. The relevant semantic predications located in the secondary analysis included:

- Coronary Artery Disease

  Diabetic Care_USES_Glucose Control

Secondary prevention_TREATS_Coronary arteriosclerosis (“Secondary prevention” referencing smoking cessation)

- Malaria

  Prophylactic treatment_USES_Amodiaquine

  Prophylactic treatment_USES_artesunate

  Prescription of prophylactic anti-malarial_USES_ Pyrimethamine

  Malaria Vaccines_TREATS_Child

  Malaria Vaccines_TREATS_Infant

As noted earlier, all reference standard interventions for acute pancreatitis were found in the primary analysis.

As predicted, in some cases in both analyses raw Semantic MEDLINE output did not precisely identify a reference standard item, but the associated citation text named the specific intervention. For example, the semantic predication “Cannulation_PREVENTS_Pancreatitis”, does not specifically name *guidewire cannulation* for acute pancreatitis; however, the associated citation text “GW [guidewire] cannulation is associated with a higher cannulation success rate and less PEP [post-ERCP pancreatitis] after pancreatic duct entry [19]” identifies the specific cannulation technique corresponding to the reference standard intervention. Nevertheless, in order for a reference standard intervention to receive true positive status, the specific intervention had to be named in the citation text. For example, there were multiple instances where “exercise” was mentioned as a preventive intervention in citations associated with the system output for coronary artery disease. Because the precise term “aerobic exercise“ did not occur, the reference standard intervention *aerobic exercise* received a false

negative status for recall assessment. To fully utilize Semantic MEDLINE with the Combo enhancement as a decision support tool, a clinician should consult the system's output of semantic predications, plus their associated citation text. An ideal interface would likely combine both, allowing the user to simultaneously review interesting semantic predications and their associated citations.

Precision and Variety of Output

The precision scores reflect the percentage of reference standard interventions included in output. However, a clinician may find the additional preventive interventions mentioned in Semantic MEDLINE's output useful. For example, the reference standard for acute pancreatitis prevention included five interventions (see Table 9). Semantic MEDLINE additionally identified antibiotic prophylaxis [20] and ulinastatin [21] as potential preventive interventions, based on the findings of randomized controlled trials. The associated DynaMed text does not discuss these potential interventions. However, other interventions in Semantic MEDLINE's output may not suit a clinical need. For example, Semantic MEDLINE also identified nafamostat mesilate [22] as a potential preventive intervention; the associated citation text notes that this intervention is "partially effective" and highlights independent risk factors associated with the disease. It is again recommended that a Semantic MEDLINE user consult the citation text (and the original article, if desired) associated with a semantic predication, to assess the relevance and strength-of-evidence pertaining to the original information need. Ideally, an interface (such as the one used by NLM) would present citation text with its associated

semantic predication, for simultaneous viewing, along with immediate access to the original PubMed record, where links to fulltext may be present.

Conclusion

Based on these findings, Semantic MEDLINE with the Combo algorithm enhancement may potentially serve as a decision support resource. It is a flexible approach to point-of-care information delivery that could be integrated into multiple environments. The authors developed the summarization software with Perl, an interpreted programming language that is compatible with multiple platforms. This Perl application provided adequate computing speeds for this project; however, to increase speed, the software could also be coded with a compiled language like Java. A locally-accessible database of SemRep output for several years' worth of MEDLINE data is also needed (for a more detailed description of how the system works, please see [7]).

Libraries could partner with the organizations they serve to customize Combo-enhanced Semantic MEDLINE for their specific user groups. For example, a library serving a healthcare organization could conduct user studies for various clientele groups to determine their information needs and preferences. The outcomes of these user studies would enable a Web designer to tailor a graphic interface for each user group. The designer could create an interface for consumers and patients, using the simplified, summarized output as a means to assist users in navigating within and understanding PubMed citation text. Another interface could assist clinicians in executing searches and accessing desired data on a single screen, organized according to their collective preferences and workflow-driven needs. Because Semantic MEDLINE, with the Combo

algorithm enhancement, is a dynamic application, users would be free to build and execute their own searches. Resources would be needed (e.g., a trained Web designer, hardware, software) to create a system customized for an institution's needs. A parent organization such as a hospital or health care system should contribute these resources if the sponsoring library cannot.

Combo-enhanced Semantic MEDLINE could either complement existing decision support products or stand alone. Because it automatically produces information relevant to multiple topics and subheading refinements, this application can potentially address the information needs of many individual users. A technician could implement the Summarization software, SemRep semantic predication database, and desired interface to freely serve clients' information needs. No subscription or licensing fees would be required. Each decision support application contributes point-of-care information in its distinctive way. Each product also has requirements enabling its practical use. Commercial products often require payment of very expensive fees, and possibly some onsite technical support. At present, Combo-enhanced Semantic MEDLINE would require substantial onsite technical support in order to establish the customized, user-centered application described in this paper. Organizations should consider their own resources and needs in choosing what value-added products they provide to their clientele.

This is an example of a technology created in part by librarians, and demonstrates a new, dynamic approach to information delivery. It surpasses the functionality of simple information retrieval, freeing users from the difficult, unrealistic task of reviewing many citations, providing instead compact summarizations of text that have been filtered for

individual information needs. This approach to information delivery could reinforce the importance of libraries as vital components in the organizations they serve.

Limitations

There are limitations in this study that warrant mention. It examined the performance of Combo algorithm-enhanced Semantic MEDLINE in terms of three disease topics, for a single subheading-type refinement. However, in an earlier study [7] the application demonstrated improved performance for a different disease topic (bladder cancer) and subheading-type refinement (genetic disease etiology) over Semantic MEDLINE with conventional, static schema summarization. Additional research is underway to examine Combo-enhanced Semantic MEDLINE's performance while processing data for additional disease topics, and an additional subheading refinement. The authors evaluated output using recommendations found in a single product (DynaMed). Similar comparisons using other commercial decision support products may shed additional light on the application's performance.

Acknowledgements

The authors express gratitude to Dr. Thomas Rindflesch and Dr. Marcelo Fiszman for their essential work in text summarization. They also wish to thank the National Library of Medicine for funding this work through grant number T15LM007123.

References

1. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc. 2005 Mar-Apr;12(2):217-24.

2. Chambliss ML, Conley J. Answering clinical questions. J Fam Pract. 1996 Aug;43(2):140-4.

3. Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J. 2006 Mar;23(1):3-12.

4. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998 Oct 21;280(15):1347-52.

5. Hoogendam A, Stalenhoef AF, Robbe PF, Overbeke AJ. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. BMC Med Inform Decis Mak. 2008;8:42.

6. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, 2004:76-83.

7. Workman TE, Hurdle JF. Dynamic summarization of bibliographic-based data. BMC Med Inform Decis Mak 2011;11:6.

8. 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 Dec;36(6):462-77.

9. Koldehoff M, Zakrzewski JL. Taurolidine is effective in the treatment of central venous catheter-related bloodstream infections in cancer patients. Int J Antimicrob Agents. 2004 Nov;24(5):491-5.

10. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

11. Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics 1951;22(1):79 – 86.

12. Riloff E. Automatically generating extraction patterns from untagged text. Proceedings of the Thirteenth National Conference on Artificial Intelligence. Menlo Park, CA: The AAAI Press/MIT Press, 1996: 1044–9.

13. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed. Inform. 2009 Oct;42(5):801-13.

14. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc. 2006:254-8.

15. Sneiderman C, Demner-Fushman D, Fiszman M, Rosemblat G, Lang FM, Norwood D, Rindflesch TC. Semantic processing to enhance retrieval of diagnosis citations from Medline. AMIA Annu Symp Proc. 2006:1104.

16. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

17. Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc. 2010 Oct;98(4):273-81.

18. Banzi R, Liberati A, Moschetti I, Tagliabue L, Moja L. A review of online evidence-based practice point-of-care information summary providers. J Med Internet Res. 2010;12(3):e26.

19. Cheung J, Tsoi KK, Quan WL, Lau JY, Sung JJ. Guidewire versus conventional contrast cannulation of the common bile duct for the prevention of post-ERCP pancreatitis: a systematic review and meta-analysis. Gastrointest Endosc. 2009 Dec;70(6):1211-9.

20. Raty S, Sand J, Pulkkinen M, Matikainen M, Nordback I. Post-ERCP pancreatitis: reduction by routine antibiotics. J Gastrointest Surg. 2001 Jul-Aug;5(4):339-45; discussion 45.

21. Tsujino T, Komatsu Y, Isayama H, Hirano K, Sasahira N, Yamamoto N, Toda N, Ito Y, Nakai Y, Tada M, Matsumura M, Yoshida H, Kawabe T, Shiratori Y, Omata M. Ulinastatin for pancreatitis after endoscopic retrograde cholangiopancreatography: a randomized, controlled trial. Clin Gastroenterol Hepatol. 2005 Apr;3(4):376-83.

22. Choi CW, Kang DH, Kim GH, Eum JS, Lee SM, Song GA, Kim DU, Kim ID, Cho M. Nafamostat mesylate in the prevention of post-ERCP pancreatitis and risk factors for post-ERCP pancreatitis. Gastrointest Endosc. 2009 Apr;69(4):e11-8.

CHAPTER 5

TEXT SUMMARIZATION AS A DECISION SUPPORT AID

T. Elizabeth Workman, M.L.I.S., Marcelo Fiszman, M.D., Ph.D., John F. Hurdle, M.D., Ph.D.

BMC Medical Informatics and Decision Making, 2011 (Submitted)

Abstract

Introduction: PubMed data potentially can provide decision support information, but PubMed is an impractical tool for that purpose. Natural language processing applications that summarize PubMed citations hold promise for unlocking that potential. **Objective**: The objective of this study was to evaluate the efficiency of a text summarization application called Semantic MEDLINE, enhanced with a novel dynamic summarization method, in identifying decision support data. **Methods**: We downloaded PubMed citations addressing the prevention and drug treatment of four disease topics. We then processed the citations with Semantic MEDLINE, enhanced with the dynamic summarization method. We also processed the citations with a conventional summarization method, as well as with a baseline procedure. We evaluated the results using clinician-vetted reference standards built from recommendations in a commercial decision support product, DynaMed. **Results**: For the drug treatment topic, Semantic MEDLINE enhanced with dynamic summarization achieved average recall and precision scores of .848 and .377, while conventional summarization produced .583 average recall and .712 average precision, and the baseline method yielded average recall and precision values of .252 and .277. In the prevention topic, Semantic MEDLINE enhanced with dynamic summarization achieved average recall and precision scores of .655 and .329. The baseline technique resulted in recall and precision scores of .269 and .247 (no conventional Semantic MEDLINE method accommodating summarization for prevention exists). Conclusion: Semantic MEDLINE with dynamic summarization outperformed conventional summarization in terms of recall, and outperformed the baseline method in

both recall and precision. This new approach to text summarization demonstrates potential in identifying decision support data for multiple needs.

Introduction

Clinicians often encounter information needs while caring for patients. Several researchers have studied this issue [1-6]. In their 2005 study, Ely and his colleagues discovered that physicians developed an average of 5.5 questions for each half-day observation, yet could not find answers to 41% of the questions for which they pursued answers [7]. Ely cites time constraints as one of the barriers preventing clinicians from finding answers. Chambliss and Conley also found that answer discovery is excessively time consuming; yet they also determined that MEDLINE data could provide answers to 71% of clinicians' questions in their separate study [8]. PubMed, the National Library of Medicine's free source for MEDLINE data, is not a practical tool for point-of-care information delivery. It generally returns excessive, often irrelevant data, even when implementing diverse search strategies [9]. Clinicians can spend an average of 30 minutes answering a question using raw MEDLINE data [10]. This is by and large due to the process of literature appraisal, which is naturally lengthened by excessive retrieval [11]. Thus this information discovery process is not practical for a busy clinical setting [10]. Applications that use natural language processing and automatic summarization of PubMed and present it in a compact form potentially can provide decision support data in a practical manner.

Objective

The objective of this study is to evaluate the performance of a new automatic summarization algorithm called Combo in identifying decision support data. To operationalize this pursuit, we incorporated the algorithm into Semantic MEDLINE, an advanced biomedical management application. We sought data on drug treatment and preventive interventions for four disease topics, and evaluated the results by comparing output to a clinician-vetted reference standard based on recommendations from a commercial decision support product, DynaMed. The Combo system was also compared to a baseline as well as a schema summarization method within the conventional Semantic MEDLINE methodology.

Background

Related Research

Natural language processing applications that summarize bibliographic text such as PubMed citations try to facilitate literature appraisal by providing succinct, relevant information suitable for point-of-care decision support. The objective of automatic text summarization is "to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's application's need" [12]. Automatic text summarization can be applied to multiple documents or information sources, [13] such as bibliographic citations retrieved from PubMed. Researchers have implemented various approaches to summarize PubMed and related data. Using an application called PERSIVAL, McKeown et. al retrieved, ranked, and summarized documents according to a patient's profile information [14]. To operate

AskHERMES, Cao and his colleagues used a machine learning approach to classify questions, and they utilized query keywords in a clustering technique for presenting output [15]. Yang and his associates clustered gene information using free text, MeSH, and Gene Ontology features, then presented summarizes based on sentence rankings [16]. Applications such as Semantic MEDLINE [17] that utilize semantic predications have the advantage of presenting a compact expression of the original information that can be filtered according to a user's specific information need. Semantic predications are succinct *subject_verb_object* declarations that simplify the meaning of the PubMed text from which they are drawn [18]. Due to their structure, they are well suited to computational analysis [19].

Semantic MEDLINE is presented to users through a Web portal that combines information retrieval, semantic processing, automatic summarization, and visualization into a single application. A user activates Semantic MEDLINE by submitting a PubMed-style keyword or MeSH query. Semantic MEDLINE's three individual components -- semantic processing (SemRep), Summarization, and Visualization -- transform MEDLINE text into concise declarations, filters these according to a user's needs, and present the results in an informative graphic display (Figure 9).

SemRep

SemRep [20] is a rule-based NLP application that interprets the meaning of abstract and title text in citations and transforms it into compact, *subject_verb_object* declarations

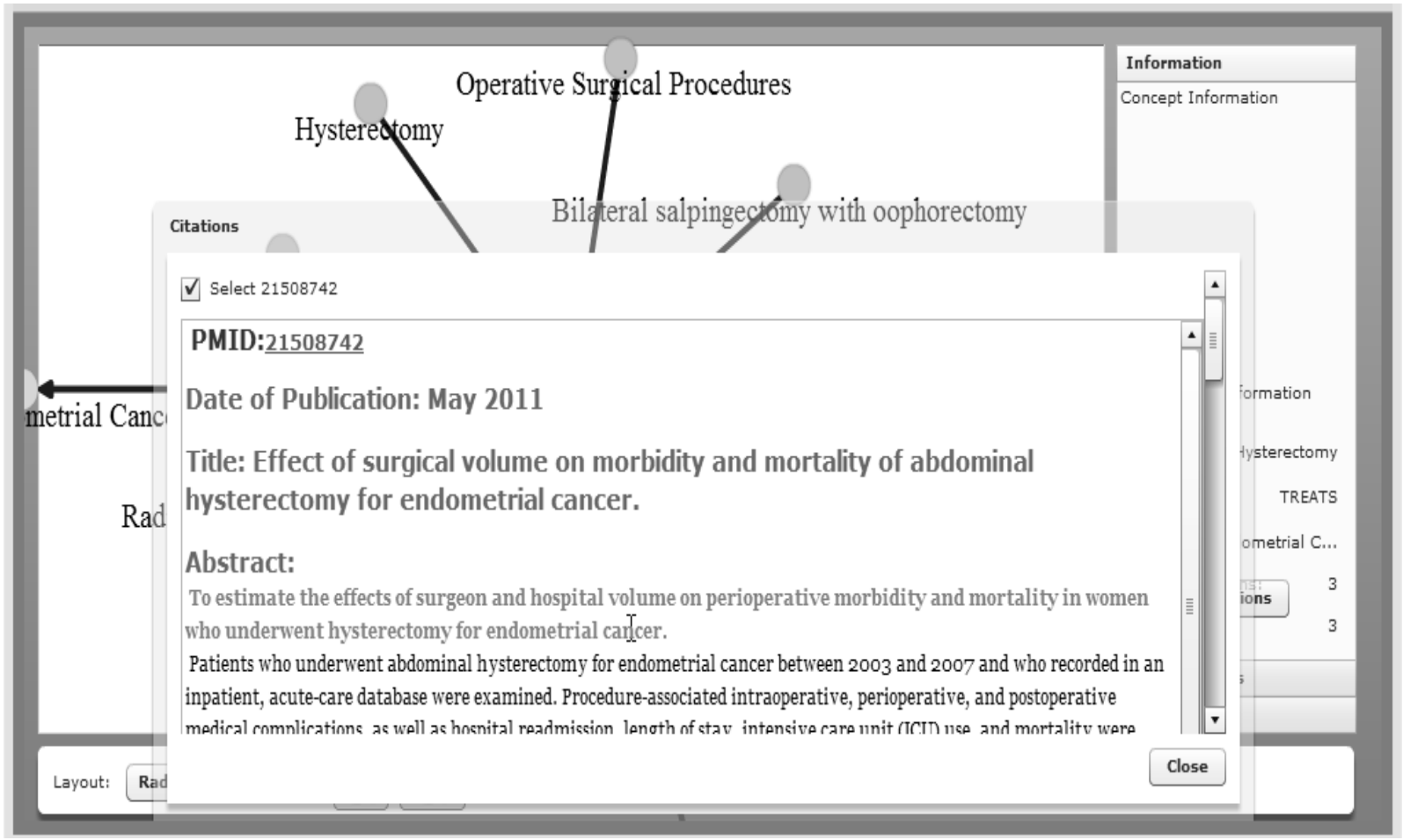

Figure 9. Semantic MEDLINE Visualized output

known as semantic predications. It draws upon resources within the Unified Medical Language System (UMLS) [21] to accomplish this. For example, if the original text is:

> "These results suggest the possibility of molecular-targeted therapy using cetuximab for endometrial cancer" [22]

SemRep produces:

```
cetuximab|phsu|TREATS|Endometrial Carcinoma|neop
```

In this example, SemRep identifies the subject and object of the original text as cetuximab and endometrial cancer, respectively. Using MetaMap [23] technology, it maps these terms to the corresponding UMLS Metathesaurus preferred concept terms *cetuximab* and *Endometrial Carcinoma,* as indicated in the resulting semantic predication. Utilizing the UMLS Semantic Network, SemRep also identifies the most likely logical semantic types associated with the subject and object, which in this case are pharmacological substance (abbreviated as *phsu*) and neoplastic process (abbreviated as *neop*). SemRep also utilizes the UMLS Semantic Network to identify the relation, or predicate, that binds the subject and object. In this case, it is *TREATS*.

Summarization

Summarization in Semantic MEDLINE [24] filters SemRep output for a "point-of-view" and a seed topic concept selected by the user. Semantic MEDLINE currently offers four points-of-view: treatment of disease; [25] substance interaction; [26] diagnosis; [27] and pharmacogenomics [28] (an application for a genetic etiology of disease point-of-view has been developed by one of us [TEW], [29] but has not yet been incorporated into the Semantic MEDLINE Web portal). For example, if the seed topic were *Endometrial*

carcinoma and the point-of-view was *treatment*, Summarization would identify semantic predications relevant to these paired concepts. Point-of-view concepts are similar to subheading refinements that can be combined with logical MeSH headings. For example, "Carcinoma, Endometrioid/*therapy*[MeSH]" could serve as a PubMed query seeking citations addressing treatment options for endometrial carcinoma. Summarization accomplishes topic and point-of-view refinements of SemRep output by subjecting it to a four-tiered sequential filter:

Relevance: Gathers semantic predications containing the user-selected seed topic. For example, if the seed topic were *Endometrial carcinoma*, this filter would collect the semantic predication cetuximab-TREATS-Endometrial carcinoma, among others.

Connectivity: Augments *Relevance* predications with those which share a nonseed argument. For example, in the above predication cetuximab-TREATS-Endometrial carcinoma, this filter would augment the *Relevance* predications with others containing *cetuximab*.

Novelty: Eliminates vague predications, such as pharmaceutical preparation-TREATS-patients, that present information that users already likely know, and are of limited use.

Saliency: Limits final output to predications that occur with adequate frequency. For example, if cetuximab-TREATS-Endometrial carcinoma occurred enough times, all occurrences would be included in the final output.

Expanding the points-of-view coverage of the Summarization filter can be done in one of two ways. The conventional approach [28] requires creating separate applications

known as schemas for each new point-of-view emphasis. This requires hard-coding specific *subject_predicate_object* patterns into the application, which limits output to predications matching the specific patterns for the new point-of-view. Prior to coding, designers must determine which patterns best capture semantic predications relevant to the given point-of-view. Conventional schema output may also be refined using degree centrality measurements [30]. The novel approach to summarization that we explore here is to produce saliency measurements on the fly, using a dynamic statistical algorithm known as Combo. [19] Combo adapts to the properties of each individual SemRep dataset by weighing term frequencies with three combined metrics. This flexibility enables summarization for multiple points-of-view, eliminates hard-coding schemas, and uses a single software application.

The Combo Algorithm to Support Summarization

The Combo algorithm combines three individual metrics to identify salient semantic predications.

Kullback-Leibler Divergence

The Kullback-Leibler divergence (KLD) [31], as applied here, assesses the values of predicates in SemRep records originating from a search query that expresses a subject paired with a point of view, (distribution P) to SemRep data with only the subject focus (distribution Q):

$$D(P||Q) = \Sigma\ P(x)\log_2(P(x)/Q(x))$$

Both distributions P and Q consist of relative frequencies for their respective predicates. Each predicate shared by each distribution receives a KLD value (before summing) indicating its value in conveying the point-of-view expressed in distribution P's search query. A database of PubMed citations from the last 10 years processed with SemRep provides the distribution Q data.

RlogF

Riloff developed the RlogF metric [32] to assess the relevance of extracted patterns consisting of a syntactic constituent (i.e., a noun or verb phrase) and their arguments (i.e., a direct or indirect object):

$$\text{RlogF}(\text{pattern}_i) = \log_2(\text{semantic type frequency}_i) * P(\text{relevant} \mid \text{pattern}_i)$$

We adapted RlogF to assess the value of a semantic type as paired with a predicate. The log of a semantic type's absolute frequency (semantic type frequency_i) is applied to the quotient of dividing that same frequency with the absolute frequency of all semantic types that are also paired with the predicate (pattern_i). We use RlogF to appraise combinations of predicates and nonseed topic semantic types. Using the example above, in `cetuximab-TREATS-Endometrial carcinoma,` the seed topic "Endometrial carcinoma" has the semantic type "neoplastic process". The opposing argument, "cetuximab", in this case has the semantic type "pharmacologic substance".

RlogF would assess the significance of “pharmacologic substance” as bound to the predicate TREATS.

PredScal

Because it assesses all predicates, KLD scores express a relative value that spans a dataset of SemRep output. RlogF scores only appraise a semantic type associated with a single predicate. Raw RlogF scores often exceed KLD scores, so we created a new metric called PredScal to scale and smooth RlogF scores according to the spatial proportions of predicates in a given SemRep dataset:

$$1 / \log_2(c)$$

Here, c represents the count of unique predicates. In rare cases where there is only one unique predicate, PredScal defaults to a value of 1.

We combine the three metrics to yield a product, which is the final Combo score:

$$KLD * RlogF * PredScal$$

Combo summarization output consists of the four highest scoring semantic $type_a$_verb_semantic $type_b$ *Relevancy* patterns (based on novel predications containing the summarization seed topic) and the four highest scoring *Connectivity* patterns (patterns sharing a nonseed topic argument’s semantic type from one of the high scoring *Relevancy* patterns).

In the Saliency phase, conventional summarization uses metrics developed by Hahn and Reimer [33] which appraise "weights" that are dependent on the predefined subject_verb_object patterns. In contrast, dynamic summarization does not utilize such predetermined patterns; instead it applies the Combo algorithm to all novel predications in order to determine which are more prominent in the data.

DynaMed

DynaMed is a decision support tool that provides intervention recommendations. In a recent study, it tied with two other products for highest ranked evidence-based decision support tool [34]. It draws upon the professional literature using a "**Systematic literature surveillance**" method in evaluating published results, using a tiered-ranking of study design types [35]. For example, here is an excerpt of the DynaMed pneumococcal pneumonia drug treatment recommendation text that we used:

Medications:

- treat for 10 days
- penicillin
    - aqueous penicillin G 600,000 units IV every 6 hours (2 million units every 4-6 hours if life-threatening)
    - procainepenicillin G 600,000 units intramuscularly every 8-12 hours
    - penicillin V 250-500 mg orally every 6 hours[36]

Methods

Disease Topics

In consultation with a clinician, we selected the four following disease topics for data acquisition:

- Arterial hypertension
- Diabetes mellitus type 2
- Congestive heart failure
- Pneumococcal pneumonia

These disease topics are significant global health concerns, and collectively have a variety of treatment options and potential preventive interventions.

Data Acquisition

We executed a single PubMed search query for each disease topic and point-of-view pairing, (i.e., drug treatment or prevention), using specific MeSH term and subheading combinations. The following lists indicate the exact MeSH terms and subheadings we used in forming these pairings:

MeSH Terms:

- hypertension
- Diabetes Mellitus, Type 2
- heart failure
- Pneumonia, Pneumococcal

Subheadings:

- drug therapy

- prevention and control

For example, to acquire citations addressing drug treatment options for pneumococcal pneumonia, we executed the search phrase "Pneumonia, Pneumococcal/*drug therapy*[Mesh]". To provide an evidence-based focus, we first restricted output to the publication types "clinical trials," "randomized controlled trials," "practice guidelines," and "meta-analyses." We then acquired citations for systematic reviews, using the publication type "review" and the keyword phrase "systematic review." Realistically, a clinician could engage Semantic MEDLINE using anything from a general keyword search to a very sophisticated search utilizing many of PubMed's search options. In addition to providing the initial topic/point-of-view pairing, this method of forming search queries also provided a middle ground within the spectrum of queries a clinician might actually use. We also restricted publication dates to coincide with the most recently published source materials DynaMed used in building their recommendations, which served as the base for our evaluative reference standards (described in detail below). We restricted the retrieval publication date in order to not retrieve materials that DynaMed curators could not have reviewed in creating their own recommendations. The eight total search queries resulted in eight separate citation datasets, each representing a pairing of one of the four disease topics with one of the two subheading concepts.

Data Processing

We processed each of the eight citation datasets separately with SemRep, then with Semantic MEDLINE utilizing the Combo algorithm. We also processed the eight SemRep output datasets with conventional Semantic MEDLINE utilizing the built-in

treatment point-of-view schema (i.e., with predetermined, hard-coded patterns). We used the following UMLS Metathesaurus preferred concepts as seed topics (required by Semantic MEDLINE) to summarize SemRep data originating from both disease/*drug treatment* and disease/*prevention and control* search query pairings:

- Hypertensive disease
- Diabetes Mellitus, Non-Insulin-Dependent
- Congestive heart failure (OR Heart Failure)
- Pneumonia, Pneumococcal

Reference Standard

We built a reference standard for each disease topic/point-of-view pairing, using vetted interventions from DynaMed, a commercial decision support product. We captured the DynaMed text for recommendations on both preventive and drug treatment interventions for each disease topic. We forwarded this text to two physician-reviewers, who highlighted the interventions they thought were viable for the associated diseases. In annotating these materials, we instructed the reviewers to ask themselves "What are the drugs used to treat this disease?" and "What interventions prevent this disease?" Disagreements between the two annotators were forwarded to a third physician adjudicator, who made the final decision regarding the conflicting annotations. The two primary reviewers were a cardiologist and a preventive medicine specialist. The adjudicator was a pathologist. We measured agreement between the two reviewers using fundamental interannotator agreement (IAA) where instances of agreement are divided by the sum of agreement instances and disagreement instances, or in other words,

matches/(matches + nonmatches). As an example, we list below the final reference standard of DynaMed arterial hypertension preventive interventions:

- Maintain normal body weight
- Reduce sodium intake
- Increased daily life activity
- Higher folate intake
- Regular aerobic physical activity
- Diet reduced in saturated and total fat
- Walking to work
- Increased plant food intake
- Diet rich in fruits, vegetables and low-fat dairy products
- Relaxation
- Whole-grain intake
- Regular tea consumption
- Limit alcohol use

The final, combined reference standards included a total of 225 interventions, with an average of approximately 28 interventions for each disease topic/point-of-view pairing. Table 13 lists the totals for all eight reference standards. The annotations of the two reviewers resulted in an average IAA score of 0.54. Table 14 lists all interannotator agreement scores.

Table 13. Reference standard intervention counts.

	Drug Treatment	**Prevention**
Arterial Hypertension	27	14
Diabetes Mellitus Type 2	55	20
Congestive Heart Failure	59	16
Pneumococcal Pneumonia	31	3

Table 14. Annotator Interrater Agreement

	Drug Treatment	Prevention
Arterial Hypertension	0.47	0.33
Diabetes Mellitus Type 2	0.73	0.44
Congestive Heart Failure	0.76	0.40
Pneumococcal Pneumonia	0.50	0.66

Baselines

We built eight baselines that simulated what a busy clinician might find when directly reviewing the PubMed citations. This is based on techniques developed Fiszman et al. [37] and Zhang et al. [30] To build baselines for the four disease topic/*drug treatment* pairings, we processed their PubMed citations with MetaMap, restricting output to UMLS Metathesaurus preferred concepts associated with the UMLS semantic group Chemicals and Drugs, and removed vague concepts using Novelty processing. Threshold values were determined by calculating the average mean of term frequencies in a baseline group, and then adding one standard deviation to the mean. In each group, all terms whose frequency scores exceeded the threshold value were retained to form the group's baseline. For example, for the congestive heart failure drug treatment group, the method extracted 1784 terms that occurred 63924 times in the MetaMap data, with a mean of approximately 35.8 occurrences per term, and a standard deviation of 154.4. This produced a cutoff threshold of 190.3. Therefore, all MetaMap terms that occurred 190 times or more were included in the congestive heart failure drug treatment baseline (a total of 72 terms).

We formed baselines for citations emerging from each disease topic/*prevention and control* pairing in a similar manner. We extracted the lines from the associated PubMed citations that contained the phrases "prevent," "prevents," "for prevention of," and "for the prevention of." These lines were processed with MetaMap, and all UMLS Metathesaurus preferred concepts associated with the UMLS disorders semantic group were removed, since the focus was preventive interventions and not the diseases themselves. Threshold values were calculated for the remaining terms, and those whose frequencies exceeded their threshold scores were retained as baseline terms.

Comparing Outputs to the Reference Standards

We evaluated outputs for the two summarization methods (Combo algorithm and conventional schema summarization) and the baselines by manually comparing them to the reference standards for the eight disease topic/subheading pairings. Since the reference standard was always a list of interventions, the comparison was straightforward. We measured recall, precision, and F_1-score (balanced equally between recall and precision).

For both summarization systems, we measured precision by grouping subject arguments by name and determining what percentage of these subject groups expressed a true positive finding. For outputs for the four disease topic/*drug intervention* pairings, we limited analysis to semantic predications in the general form of "Intervention X_TREATS_disease Y", where the object argument reflected the associated disease concept. If the subject intervention X argument matched a reference standard intervention, that intervention received a true positive status. In similar predications

where the subject argument was a general term, such as “intervention regimes”, we examined the original section of citation text associated with the semantic predication. If this citation text indicated a reference standard intervention it received a true positive status. For example, in the dynamic summarization output for arterial hypertension prevention, the semantic predication “Dietary Modification_PREVENTS_Hypertensive disease” summarized citation text that included advice for dietary sodium reduction [38]; therefore, the reference standard intervention “reduce sodium intake” received a true positive status.

Only the combo algorithm summarized output for the four disease topic/*prevention and control* pairings was compared to the reference standard, since there is no conventional schema for prevention. In addition to predications in the form “Intervention X_PREVENTS_disease_Y,” other predications where argument concepts had prevention terms such as “Exercise, aerobic_AFFECTS_blood pressure” and “Primary Prevention_USES_Metformin” were used.

We evaluated each baseline by comparing its terms to those of its associated reference standard. If a term in a baseline matched an intervention in the relevant reference standard, the baseline term received a true positive status. We also assigned true positive status to less specific baseline terms if they could logically be associated with related reference standard interventions. For example, in the baseline for pneumococcal pneumonia prevention the term “Polyvalent pneumococcal vaccine” was counted as a true positive, even though it did not identify a specific polyvalent pneumococcal vaccine that was on the reference standard.

Results

The PubMed search queries retrieved varying quantities of output, as did SemRep, conventional, and dynamic summarization. Table 15 lists PubMed output citation quantities according to disease topic and point-of-view. Tables 16 - 18 list quantitative outputs for the other processes.

System Performance

Performance metric outcomes are listed in Tables 19 - 20. No conventional schema is available in summarizing for a prevention point-of-view; therefore, just the Combo algorithm enhanced summarization and the baseline method performance outcomes are included.

Table 15. Citation retrieval results

	Drug Treatment	**Prevention**
Arterial Hypertension	12335	875
Diabetes Mellitus Type 2	3716	435
Congestive Heart Failure	3256	344
Pneumococcal Pneumonia	115	81

Table 16. SemRep semantic predication outputs

	Drug Treatment	Prevention
Arterial Hypertension	94353	4836
Diabetes Mellitus Type 2	37962	2654
Congestive Heart Failure	28951	2630
Pneumococcal Pneumonia	918	643

Table 17. Combo algorithm-enhanced summarization semantic predication output

	Drug Treatment	Prevention
Arterial Hypertension	13015	279
Diabetes Mellitus Type 2	3237	188
Congestive Heart Failure	4175	207
Pneumococcal Pneumonia	189	137

Table 18. Conventional treatment schema semantic predications output

	Drug Treatment
Arterial Hypertension	8052
Diabetes Mellitus Type 2	2645
Congestive Heart Failure	2375
Pneumococcal Pneumonia	62

Table 19. Performance Metrics, Drug Treatment Point-of-View, for Combo-enhanced dynamic summarization (DS), conventional treatment schema (TS), and baseline (BL) methodologies**.**

Disease	**Recall**			**Precision**			**F_1-Score**		
	DS	TS	BL	DS	TS	BL	DS	TS	BL
Arterial Hypertension	0.93	0.82	0.26	0.39	0.73	0.41	0.55	0.77	0.32
Diabetes Mellitus Type 2	0.89	0.56	0.35	0.35	0.68	0.25	0.50	0.62	0.29
Congestive Heart Failure	0.93	0.70	0.13	0.34	0.60	0.25	0.50	0.64	0.17
Pneumococcal Pneumonia	0.65	0.26	0.19	0.43	0.83	0.32	0.51	0.39	0.24

Table 20. Performance Metrics, Prevention Point-of-View, for Combo-enhanced dynamic summarization (DS), and baseline (BL) methodologies.

Disease	**Recall**		**Precision**		**F_1-Score**	
	DS	BL	DS	BL	DS	BL
Arterial Hypertension	0.77	0.23	0.13	0.13	0.22	0.17
Diabetes Mellitus Type 2	0.68	0.18	0.50	0.33	0.58	0.24
Congestive Heart Failure	0.50	0.33	0.30	0.31	0.37	0.32
Pneumococcal Pneumonia	0.67	0.33	0.39	0.22	0.49	0.26

Discussion

The evaluation results imply that dynamic text summarization with the Combo algorithm provides a viable alternative to direct review of PubMed citations for locating decision support data. This is encouraging, because dynamic summarization could expand the value of Semantic MEDLINE at the point-of-care. Performance improvements over the baseline methodology can be seen in both recall and precision results. Including findings from both drug treatment and prevention analyses, Combo produced average recall and precision scores of 0.75 and 0.35, while the baseline method yielded average recall and precision values of 0.25 and 0.28. Combo summarization outperformed the baseline methodology by an average F_1-score margin of 0.21. The Combo algorithm especially performed well in terms of recall for large datasets. For the three disease topic/point-of-view pairings whose initial citation input exceeded 1000 (the drug treatment topics of arterial hypertension, diabetes mellitus type 2, and congestive heart failure) average recall was 0.916.

Drug Treatment Results

Combo algorithm-enhanced dynamic summarization outperformed conventional summarization and the baseline method in recall, but was outperformed by conventional summarization in terms of precision. Combo summarization achieved 0.85 average recall, and 0.38 average precision. The conventional schema produced average recall and precision scores of 0.59 and 0.71. Both dynamic summarization and conventional summarization outperformed the baseline method, which produced average recall and precision scores of 0.23 and 0.31. Based on these findings, if a clinical user wished to

locate the maximum amount of drug treatment options using one of these three methods, Combo would be the better choice. On the other hand, the new method is less precise, but this effect is moderated by the visualization tool that Semantic MEDLINE offers. Visualization conveniently presents all citation data (including the text of the abstract itself) that are relevant to an Intervention X_TREATS_disease Y relationship in an easily viewed, reader-friendly display. Viewed in context, clinicians can quickly discard irrelevant treatments. We would argue that recall is more critical in clinical browsing than precision. The cognitive load required to dismiss a false positive is lower than trying to deduce a missing (false negative) treatment. We chose to use the standard F_1-score because it is more conventional, but if we weight recall more, in line with the argument above, then the Combo summarization would be quite competitive with the conventional technique.

Prevention Results

Combo summarization was less effective in identifying preventive interventions in the relevant reference standards, producing an average recall of 0.66 and an average precision rate of 0.33. There are two obvious possibilities for this diminished efficiency. First, the citation sets were substantially smaller than three of the four drug treatment citation sets, thus providing less initial data. As with most statistical techniques, larger sample sizes tend to lead to better performance. Second, preventive interventions described in text are often more general than drug therapies. For example, “lifestyle changes” may be more difficult to interpret in the SemRep phase. Also, the lower interannotator agreement scores suggest that clinicians are less apt to agree on prevention

standards. This may also be reflected in the professional literature. Dynamic summarization with the Combo algorithm outperformed the baseline methodology, which produced an average recall of 0.27 and an average precision of 0.25. This suggests that dynamic summarization is a superior alternative to directly reviewing PubMed citations for identifying preventive interventions.

Error Analysis

We classified false positive findings by type, and false negative findings by the first sequential data source (i.e., PubMed, SemRep output, Dynamic Summarization output) that did not include them.

False Positives

Most of the false positives for both drug treatment and prevention points-of-view could be classified as unproductive general subject arguments; pharmaceuticals or supplements not included in the relevant reference standards; or other therapies not included in the relevant reference standards. In the prevention data, pharmaceuticals or supplements not included in the relevant reference standards accounted for 62.5% of all false positives, while unproductive general subject arguments and other therapies not included in the relevant reference standards accounted for 17.5% and 15.5%, respectively. In the drug treatment data, pharmaceuticals or supplements not included in the relevant reference standard accounted for an even greater percentage of false positives at 73.7%, while unproductive general subject arguments and other therapies not included in the relevant reference standard accounted for 14.2% and 12%.

There are multiple possible reasons why there was such a high percentage of nonreference standard pharmaceutical or supplement false positives. Initial citation retrieval was not limited by a beginning publication date. In other words, all search queries retrieved relevant citations for as far back in time as PubMed makes available. Therefore, information retrieval likely included older drugs which had been replaced by newer medications as preferred treatments. Also, we used a single data source in creating the reference standard. If we had included recommendations from other decision support tools in addition to those from DynaMed, the final reference standard might have included other treatments found within this false positive classification. Another data trend potentially contributed to reduced precision. Subject arguments that occurred two times or less in an output for a given disease topic/point-of-view pairing accounted for 69.7% of all false positives. If all such results were removed from the data, precision would increase, with a proportionately small effect on recall.

False Negatives

Because Semantic MEDLINE is a pipeline application, data loss can be tracked by documenting the first sequential process (among PubMed retrieval, SemRep, and Dynamic Summarization) that does not include a reference standard intervention. We applied this method in analyzing false negative interventions to determine which process "lost" the desired data. In tracking the 23 false negatives that addressed a drug treatment point-of-view, PubMed retrieval did not garner 43.5% (10 false negatives); SemRep output did not include 47.8% (11 false negatives); and dynamic summarization did not identify 8.7% (2 false negatives). False negatives emerging from the prevention point-

of-view data were slighted more balanced. In this case, PubMed retrieval did not include 41.2% (7 false negatives) while SemRep output did not include 35.3% (6 false negatives) and dynamic summarization output did not include 23.5% (4 false negatives). However, in analyses for both points-of-view, dynamic summarization performed better than the other two processes. Visualization output was not included; it was considered irrelevant, since it automatically includes all output from summarization.

PubMed Retrieval Volume and Performance

Performance measurements suggest a system preference for larger citation input. Among search queries pairing the disease topics with the drug therapy subheading, the only query resulting in a relatively small amount of citations (the pneumonia pneumococcal query) also lead to comparatively diminished performance. System performance for pneumococcal pneumonia drug treatment data produced only 0.65 recall, while the other disease topic/drug treatment pairings achieved 0.89 or higher recall. System performance for prevention had similar results, with recall ranging from 0.50 to 0.76, with overall fewer citations than the drug treatment data. However, in a pilot project the system produced 100% recall for prevention data on a single disease topic (acute pancreatitis), with only 156 citations [39]. We conclude that citation volume can be a factor for some clinical topics, but not for all of them. In cases like acute pancreatitis, where therapeutic options are narrow, the system can summarize successfully despite a relatively sparse citation set.

Limitations

There are limitations in this study. It explores summarization for only two points-of-view (prevention and drug treatment) for the single task of decision support. However, an earlier study examined Combo-enhanced dynamic summarization for a genetic disease etiology point-of-view, within the task of secondary genetic database curation [19]. The curation study revealed improved summarization performance for that task. In this current study, we examined dynamic summarization for just four disease topics. However, a pilot project [39] featuring three different disease topics (acute pancreatitis, coronary artery disease, and malaria), again within the context of preventive intervention decision support, produced slightly superior results. This creates optimism that this text summarization method may enable others to locate decision support data. Finally, we evaluated system output with recommendations garnered from a single commercial decision support product. Comparing performance to other decision support sources may shed further light on Combo-enhanced dynamic summarization as a potential decision support tool.

Conclusion

In order to evaluate the performance of a new dynamic text summarization extension (Combo) to Semantic MEDLINE, we applied it, plus conventional Semantic MEDLINE, and a baseline summarization methodology (designed to mimic manual clinical review) to a clinical decision support task. We chose four disease topics and processed PubMed citations addressing their drug treatment and prevention. We processed the citations with SemRep, an application that transforms PubMed text into semantic predications. We

processed the SemRep output using the three summarization methodologies. An evaluation using reference standards (clinically vetted DynaMed) showed that the new summarization method outperformed the conventional application and baseline methodology in terms of recall, while the conventional application produced the highest precision. Dynamic and conventional summarization were superior to the baseline methodology. These findings imply that the new text summarization application holds potential in assisting clinicians in locating decision support information.

Abbreviations

NLP: Natural Language Processing; UMLS: Unified Medical Language System

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

TEW designed the study; downloaded the citations; oversaw data processing with SemRep, the conventional treatment summarization schema, and the dynamic summarization application; built the baseline measurements; coordinated reference standard construction; performed the manual data evaluation; and wrote the original manuscript. MF guided the data evaluation and provided essential manuscript revisions. JFH contributed to the Combo algorithm design by suggesting use of the RlogF metric; provided guidance in the reference standard vetting process, and also provided essential manuscript revisions.

Acknowledgements and Funding

We express our gratitude to Denise Beaudoin, Bruce Bray, and Stan Huff for serving as data reviewers. We also thank Stephane Meystre for his counsel in disease topic selection. We also thank the National Library of Medicine for funding this work through grant number T15LM007123.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med. 1985 Oct;103(4):596-9.

2. Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Med Decis Making. 1995 Apr-Jun;15(2):113-9.

3. Alper BS, Stevermer JJ, White DS, Ewigman BG. Answering family physicians' clinical questions using electronic medical databases. J Fam Pract. 2001 Nov;50(11):960-5.

4. Bergus GR, Randall CS, Sinift SD, Rosenthal DM. Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues? Arch Fam Med. 2000 Jun;9(6):541-7.

5. Graber MA, Randles BD, Monahan J, Ely JW, Jennissen C, Peters B, Anderson D. What questions about patient care do physicians have during and after patient contact in the ED? The taxonomy of gaps in physician knowledge. Emerg Med J. 2007 Oct;24(10):703-6.

6. Graber MA, Randles BD, Ely JW, Monnahan J. Answering clinical questions in the ED. Am J Emerg Med. 2008 Feb;26(2):144-7.

7. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. J Am Med Inform Assoc. 2005 Mar-Apr;12(2):217-24.

8. Chambliss ML, Conley J. Answering clinical questions. J Fam Pract. 1996 Aug;43(2):140-4.

9. Golder S, McIntosh HM, Duffy S, Glanville J. Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE. Health Info Libr J. 2006 Mar;23(1):3-12.

10. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998 Oct 21;280(15):1347-52.

11. Hoogendam A, Stalenhoef AF, Robbe PF, Overbeke AJ. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital. BMC Med Inform Decis Mak. 2008;8:42.

12. Mani I. Automatic summarization. Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2001.

13. Hahn U, Mani I. The challenges of automatic summarization. Computer 2000;33(11):29 - 36.

14. McKeown K, Elhadad N, Hatzivassiloglou V. Leveraging a common representation for personalized search and summarization in a medical digital library. 3rd ACM/IEEE-CS joint conference on Digital libraries. Houston, TX, 2003: 159–70.

15. Cao Y, Liu F, Simpson P, Antieau L, Bennett A, Cimino JJ, Ely J, Yu H. AskHERMES: An online question answering system for complex clinical questions. J Biomed Inform. 2011 Apr;44(2):277-88.

16. Yang J, Cohen AM, Hersh W. Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts. AMIA Annu Symp Proc. 2007:831-5.

17. Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch T. Semantic MEDLINE: a web application to manage the results of PubMed searches. . Proceedings of the Third International Symposium on Semantic Mining in Biomedicine, 2008: 69 -76.

18. 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 Dec;36(6):462-77.

19. Workman TE, Hurdle JF. Dynamic summarization of bibliographic-based data. BMC Med Inform Decis Mak. 2011;11:6.

20. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller S, Hersh W, Friedman C, eds. Medical informatics: knowledge management and data mining in biomedicine. New York: Springer, 2005: 399-422.

21. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

22. Takahashi K, Saga Y, Mizukami H, Takei Y, Machida S, Fujiwara H, Ozawa K, Suzuki M. Cetuximab inhibits growth, peritoneal dissemination, and lymph node and lung metastasis of endometrial cancer, and prolongs host survival. Int J Oncol. 2009 Oct;35(4):725-9.

23. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

24. Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics, 2004:76-83.

25. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.

26. Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in Medline citations. AMIA Annu Symp Proc. 2006:254-8.

27. Sneiderman C, Demner-Fushman D, Fiszman M, Rosemblat G, Lang FM, Norwood D, Rindflesch TC. Semantic processing to enhance retrieval of diagnosis citations from Medline. AMIA Annu Symp Proc. 2006:1104.

28. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

29. Workman TE, Fiszman M, Hurdle JF, Rindflesch TC. Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information. J Med Libr Assoc. 2010;98(4):273 - 81.

30. Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform. 2011 May 8.

31. Kullback S, Leibler RA. On information and sufficiency. Annals of Mathematical Statistics 1951;22(1):79 – 86.

32. Riloff E. Automatically generating extraction patterns from untagged text. Proceedings of the Thirteenth National Conference on Artificial Intelligence. Menlo Park, CA: The AAAI Press/MIT Press, 1996: 1044–9.

33. Mani I, Maybury MT. Advances in automatic text summarization. Cambridge, Mass.: MIT Press, 1999.

34. Banzi R, Liberati A, Moschetti I, Tagliabue L, Moja L. A review of online evidence-based practice point-of-care information summary providers. J Med Internet Res. 2010;12(3):e26.

35. Content/Editorial Policies: EBSCO Publishing, 2010. [2010; cited 16 August 2011]. <http://www.ebscohost.com/dynamed/content.php>.

36. DynaMed. Pneumococcal Pneumonia: EBSCO Industries, Inc, 2011. [2011].

37. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.

38. Khan NA, Hemmelgarn B, Padwal R, Larochelle P, Mahon JL, Lewanczuk RZ, McAlister FA, Rabkin SW, Hill MD, Feldman RD, Schiffrin EL, Campbell NR, Logan AG, Arnold M, Moe G, Campbell TS, Milot A, Stone JA, Jones C, Leiter LA, Ogilvie RI, Herman RJ, Hamet P, Fodor G, Carruthers G, Culleton B, Burns KD, Ruzicka M, deChamplain J, Pylypchuk G, Gledhill N, Petrella R, Boulanger JM, Trudeau L, Hegele RA, Woo V, McFarlane P, Touyz RM, Tobe SW. The 2007 Canadian Hypertension Education Program recommendations for the management of hypertension: part 2 - therapy. Can J Cardiol. 2007 May 15;23(7):539-50.

39. Workman TE, Stoddart JM. Rethinking information delivery: using a natural language processing application for point-of-care data discovery. J Med Libr Assoc. Forthcoming 2011.

CHAPTER 6

CONCLUSION

Summary

Hypothesis Validation

As noted in Chapter 1, the central and subhypotheses can be operationalized and validated through simulating human tasks. The work of Chapters 2 -5 operationalized efforts to test the central hypothesis that an NLP text summarization process that transforms bibliographic text into a topically filtered, compact form can be used to extract and identify data crucial to multiple information needs. The work of Chapters 3 – 5 operationalized efforts exploring the subhypothesis that once it is transformed into a basic compact form, bibliographic text collectively retains the thematic focus that was expressed in the initial search query used to retrieve it. This was demonstrated by using the Combo algorithm as the *Saliency* filtering mechanism (within the four-tier summarization framework) to summarize PubMed text focused on multiple disease topic/point-of-view pairings, for multiple tasks. The three tasks:

- Secondary genetic database curation
- Identifying clinical decision support for preventive intervention discovery
- Identifying clinical decision support for drug treatment intervention discovery

garnered positive results concerning the central hypothesis. The conventional genetic disease etiology summarization software provided data whose curation appeal had previously been confirmed, because it was featured in the Genetics Home Reference, Online Mendelian Inheritance in Man, and Entrez Gene databases. The conventional treatment summarization software also provided verified data in the form of intervention recommendations from a respected decision support product. The Combo algorithm, through manual calculation within the Summarization four-tier filtering framework, also

provided validated data for secondary genetic database curation, without explicit point-of-view constraint, using data originating from a PubMed search which expressed a desired disease topic/point-of-view thematic focus. The Combo algorithm-enhanced software also facilitated dynamic summarization by identifying relevant drug treatment and preventive interventions, again without explicit point-of-view constraint, using data originating from PubMed search queries expressing the desired disease topics/points-of-view thematic focuses. The Combo algorithm output validated the subhypothesis.

Individual Performance Metrics

While the work of the three aims resulted in successful information identification, the success of completing the three tasks varied. The conventional genetic disease etiology summarization software located a total of six gold standard genes in the first curation study, and three in the second. Dynamic summarization located eight gold standard genes in the second curation study (it was not applied in the first study). In the second study, dynamic summarization produced 0.61 recall, 0.81 precision, and an F-score of 0.69. Conventional summarization produced 0.23 recall, 1.0 precision, and an F-score of 0.37. In the pilot prevention study, after performing the primary and secondary analyses, dynamic summarization achieved an average recall of 0.79, average precision of 0.45, and an average F-score of 0.57. In the larger study documented in Chapter 5, dynamic summarization achieved 0.656 average recall, 0.329 average precision, and an average F-score of 0.41. In drug treatment data performance, conventional summarization produced 0.583 average recall, 0.712 average precision and an average F-score of 0.60. Dynamic summarization produced 0.848 average recall, 0.377 average precision, and an average F-

score of 0.51. Dynamic and conventional summarization outperformed the baseline metric where it was applied.

Discrepancies in dynamic summarization's performance for the two preventive intervention studies have several possible explanations. In the first study, we evaluated general subject arguments for reference standard items by looking at the entire text associated with a predication; in the second study, we only examined the span of text a given semantic predication summarized. These different study designs are intentional. The first preventive intervention study was done on behalf of the Medical Library Association, and our goal for this study was in part to duplicate a user's actions in reviewing all associated citation data (because this is the way Visualization presents citation data to a user when a desired arc is clicked). In the second study, we simply observed the connection between the more general semantic predications and the exact text they summarized. In the first study, we only looked at predications in the form "Intervention X_PREVENTS _disease Y" to calculate precision. In the second study, we reviewed all predications to calculate the value. Finally, dynamic summarization achieved 1.0 recall for one of the three diseases in the first study; while this is exciting, it may have also in a sense skewed the results.

Variables Within the Semantic MEDLINE Model

There are other variables within the Semantic MEDLINE model, (outside of Summarization processing), that effect Summarization performance. I used MeSH search strategies for all information retrieval operations in all four separate studies. The search query used in Chapter 2 (evaluating the conventional genetic disease etiology

summarization software) implemented a single MeSH term with keyword phrases. All search queries used to evaluate Combo-enhanced summarization implemented MeSH terms combined with subheadings, and short keyword phrases when needed. While searching with MeSH terms leverages the work of expert indexers, MeSH term representation within PubMed is not consistent, nor does it completely capture biomedical concepts [1-3]. While employing simple MeSH term/subheading searches enabled validation of the work's subhypothesis, future search queries that are more complex may result in improved information retrieval [4].

Due to various limitations, SemRep does not capture all possible semantic predications in citation text. SemRep relies on the UMLS Metathesaurus, which does not completely provide correlating terms to multiple types of external terminologies [5-7]. SemRep does not accommodate all 54 predications [8] in the UMLS Semantic Network [9], which in turn contains inaccuracies [10].

Reference Standard Limitations

The reference standards used in Chapters 3 – 5 should not be considered definitive lists of best interventions for their associated diseases and points-of-view. DynaMed served as the exclusive electronic source of these interventions; however, there are other well-received decision support products. In Banzi's study [11], DynaMed was one of three products tied as the top ranked resource. The other two were EBM Guidelines [12] and UpToDate [13]. The University of Utah Health Sciences Center does not subscribe to EBM Guidelines. DynaMed's presentation format was superior to that of UpToDate, in terms of presenting data to reviewers. For these reasons (accessibility, ranking, and

format presentation) I chose DynaMed as the reference standards' source. However, by Banzi's own disclosure, no product established itself as "the best" when judged with the criteria used in his study. It is quite possible that other decision support products may have suggested other interventions.

Significance of This Work to the Field of Biomedical Informatics

The research in this dissertation points to means in which Summarization may benefit many stakeholders in the health science community. Genetic database curators may save much time and effort by subject PubMed search results to Semantic MEDLINE process, instead of directly reviewing the citations. Clinicians may have faster access to decision support data and be spared the experience of reviewing large datasets of PubMed retrieval. Semantic MEDLINE also holds potential in assisting health consumers to navigate within PubMed retrieval, and understand passages of complicated citation text. Because dynamic summarization has a demonstrated capacity to summarize text for multiple points-of-view, users can use it in pursuing many different information needs.

Mechanisms and artifacts in this research have also provided new methodologies that may help other biomedical informatics researchers. The work of Chapter 5 included development of a novel evaluative technique. To evaluate the results for prevention, I automatically extracted sections of citation text containing phrases such as "prevents" and "for the prevention of" and then processed these sections with MetaMap to find preventive data found in the citation text. I also combined techniques developed by Fiszman [14] and Zhang [15] in order to combine an established method of locating interventions in citation text along with an automated threshold cutoff calculation to form

an evaluative baseline. These techniques may help other scientists in evaluating NLP output or citation analysis in a biomedical informatics context.

Future Directions

Potential users may benefit from interfaces specifically designed for text summarization. A new branch of research focused on text summarization within the human/computer interaction paradigm may result in new information systems providing new levels of service to their users. A critical component in this research pursuit would be the study of information seeking behavior within text summarization applications. Many researchers have laid a foundation in information seeking behavior theory. [16] There are several individual schools of thought in this domain. Theories which model the role of affect in user behavior hold potential in leading to text summarization system interface design. For example, the IRU methodology [17] of modeling user behavior, designed by Dr. Diane Nahl, documents [18] the role of affect, in addition to cognition and sensorimotor function, in human/computer interactions. Techniques like Nahl's enable observers to model information seeking behavior, and provide cues for developers to design more efficient, user-friendly systems. Affect is important and influential, especially in scenarios of stress, where a patient may be seeking information on a catastrophic disease, or a clinician may be seeking elusive but crucial information to save a patient.

There are potential new uses for Combo-enhanced Semantic MEDLINE text summarization. Because it can distill and organize large volumes of data, it may serve in knowledge discovery, for example, leading scientists to undiscovered connections

between diseases and treatments, or disease and genetic mutations. It may also assist universities, health organizations and institutes that fund research in identifying areas where there is a need for more research, where the scientific community should concentrate new efforts. Fulltext clinical guidelines could also be processed by Combo-driven Semantic MEDLINE in order to find data prevalent to a specific health issue, expressed as a UMLS Metathesaurus seed concept. In order to successfully do this, SemRep would need modifications to enable it to process fulltext items. Combo-driven Semantic MEDLINE summarization could also be integrated into an electronic medical record (EMR) environment, where it could automatically provide information by utilizing coded values. This service could provide information originating from PubMed citations, much like MedlinePlus Connect [19] provides consumer-oriented health information by drawing on the MedlinePlus pool of consumer resources. Following the MedlinePlus Connect system, such a service could easily provide drug information (using RxNorm [20] or NDC coding [21]), information regarding lab tests (using LOINC [22]), or diagnoses (using ICD-9-CM coding [23]). All of the associated codes could be mapped to corresponding MeSH terms within the UMLS Metathesaurus, which then would be combined with the appropriate subheading to form a PubMed query. The resulting citations would be processed by SemRep. The associated EMR code(s) would be mapped to the closest corresponding Metathesaurus preferred concept. This preferred concept would serve as the seed topic in dynamic summarization. Summarized results would be displayed in an interactive graph, where users could select the semantic predications addressing their information needs.

Abstracting the Model

The skeletal framework of the Semantic MEDLINE model pursued in this work can potentially be applied to other data types, such as clinical text and Internet data, and other kinds of tasks. For example, the essential pipeline functions of retrieving data with a focused search, transforming it into subject_verb_object triplets, and summarizing results using a seed topic concept can potentially be applied to clinical text. This potential can be considered sequentially by process. A researcher wishes to find prominent issues expressed in electronic health data for patients diagnosed with fibromyalgia. She retrieves all electronic medical records (EMRs) containing the keyword "fibromyalgia" and ICD-9-CM code 729.1, *Myalgia and myositis, unspecified*, the code associated with fibromyalgia, according to 2011 CDC documentation [24].

The EMR text is transformed in subject_verb_object triplets using the next application in the pipeline. SemRep is designed to handle bibliographic citation data. Another application may perform better than SemRep. Recent research provides potential direction for developing such an application. In response to the 2010 i2b2 challenge [25], de Bruijn and colleagues [26] developed three machine learning applications that together identify and classify medical issues in clinical text, and then identify the relationships that bind the concepts. Their system utilizes MEDLINE records and UMLS output, in conjunction with cTAKES [27] and MetaMap [28] technologies, to complete the three tasks. This pipeline application performed well in the 2010 i2b2 challenge (first place in the concept identification and classifications tasks; second place in the relation identification task) and could possibly produce triplets for the fibromyalgia task. Rink and his colleagues [29] developed a supervised machine learning application

that outperformed de Bruijn's relation identification application. Rink's application utilizes WordNet [30] and the General Inquirer lexicon [31] as external knowledge sources. Replacing de Bruijn's relation extraction method with Rink's would likely improve the pipeline's performance. An alternative triplet producing resource might function like a system developed by Khoury and his colleagues [32]. Khoury's application only requires an unannotated training corpus of domain-specific text. It then utilizes a fuzzy-logic statistical algorithm to identify concepts and their binding relationships. Khoury's approach assumes a subject-verb-object concept representation in the original text, so it might not detect certain kinds of information, such as that written in passive voice.

The triplet data would then be summarized using Combo-driven software. This software could initially present a list of subject and object arguments, organized in descending order by frequency, from which the user would choose a summarization seed topic. To build the Q distribution for Kullback-Leibler Divergence computations, the system could build an approximate profile of similar data by retrieving EMR records containing all ICD-9-CM codes 710 – 739, representing Diseases of the Musculoskeletal System and Connective Tissue and transforming it into subject_verb_object triplets.

The essential pipeline functions of Combo-driven Semantic MEDLINE could also be applied to Internet data. Assume that a public health officer wants to monitor current flu epidemic information as reported in Web-based news stories. He could harvest Internet articles using Google's news utility [33] using a search such as "flu season" or "flu cases" and limit recall to articles from the last 30 days. These articles could be converted into subject_verb_object triplets using an all-purpose application such as Khoury's. The

triplets could be summarized with the Combo-driven software, using a similar approach as the one described for clinical text. In order to build the Q distribution for Kullback-Leibler Divergence computations, a more general Google News search such as “disease epidemic” could yield appropriate articles for building the Q distribution.

References

1. Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983 Apr;71(2):176-83.

2. Murphy LS, Reinsch S, Najm WI, Dickerson VM, Seffinger MA, Adams A, Mishra SI. Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexers and investigators. BMC Complement Altern Med. 2003 Jul 7;3:3.

3. Portaluppi F. Consistency and accuracy of the Medical Subject Headings thesaurus for electronic indexing and retrieval of chronobiologic references. Chronobiol Int. 2007;24(6):1213-29.

4. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Stud Health Technol Inform. 2004;107(Pt 1):311-6.

5. Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. J Am Med Inform Assoc. 2010 Nov-Dec;17(6):675-80.

6. Boxwala AA, Zeng QT, Chamberas A, Sato L, Dierks M. Coverage of patient safety terms in the UMLS metathesaurus. AMIA Annu Symp. Proc 2003:110-4.

7. Keselman A, Smith CA, Divita G, Kim H, Browne AC, Leroy G, Zeng-Treitler Q. Consumer health concepts that do not map to the UMLS: where do they fit? J Am Med Inform Assoc. 2008 Jul-Aug;15(4):496-505.

8. Semantic network. UMLS Reference Manual. Bethesda, MD: National Library of Medicine (US).

9. Rindflesch TC. SemRep predicates. In: A pdf document describing the predicates that SemRep accommodates ed. Salt Lake City, UT, 2011.

10. Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform. 2003 Dec;36(6):450-61.

11. Banzi R, Liberati A, Moschetti I, Tagliabue L, Moja L. A review of online evidence-based practice point-of-care information summary providers. J Med Internet Res. 2010;12(3):e26.

12. EBM Guidelines. John Wiley and Sons, Inc.

13. UpToDate. UpToDate, Inc.

14. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.

15. Zhang H, Fiszman M, Shin D, Miller CM, Rosemblat G, Rindflesch TC. Degree centrality for semantic abstraction summarization of therapeutic studies. J Biomed Inform. 2011 May 8.

16. Fisher K, Erdelez S, McKechnie L, eds. Theories of information behavior. Medford, NJ: Information Today, 2005.

17. Nahl D. Social–biological information technology: an integrated conceptual framework. J Am Soc Inf Sci Technol. 2007 58(13):2021–46.

18. Nahl D. A discourse analysis technique for charting the flow of micro-information behavior. Journal of documentation 2007;63(3):323 - 39.

19. MedlinePlus Connect: National Library of Medicine, 2011. [rev. 16 June 2011 2011; cited 19 Sept 2011]. <http://www.nlm.nih.gov/medlineplus/connect/overview.html>.

20. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011 Jul-Aug;18(4):441-8.

21. National Drug Code directory. 2011 ed: Food and Drug Administration; Health and Human Services, 2011.

22. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem. 2003 Apr;49(4):624-33.

23. ICD-9 provider and diagnostic codes, overview: Centers for Medicare and Medicaid Services; Health and Human Services. [rev. 11 July 2011; cited 19 Sept 2011]. <http://www.cms.gov/ICD9ProviderDiagnosticCodes/01_overview.asp>.

24. International classification of diseases, ninth revision, clinical modification (ICD-9-CM); Index of ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2011/: Centers for Disease Control and Prevention. [cited 30 September 2011]. <ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2011/>.

25. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):552-6.

26. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):557-62.

27. Savova G, Kipper-Schuler K, Buntrock J, Chute C. UIMA-based clinical information extraction system. Proceedings of The Sixth International Conference on Language Resources and Evaluation: Workshop on Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP Marrakech, Morocco: Language Resources and Evaluation Conference, 2008: 39 - 42.

28. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.

29. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):594-600.

30. Fellbaum C. WordNet: an electronic lexical database. Cambridge, MA: MIT Press, 1998.

31. Stone PJ, D.C. D, Smith MS. The General Inquirer: a computer approach to content analysis. Cambridge, MA: MIT Press, 1966.

32. Khoury R, Karray F, Yu Sun Y, Kamel M, Basir O. Semantic understanding of general linguistic items by means of fuzzy set theory. IEEE Transactions on Fuzzy Systems 2007;15(5):757 - 71.

33. Google News. [rev. 3 October 2011]. <http://news.google.com>.