Extracting Patient Profiles from Patient Records and Online Literature

Vasileios Hatzivassiloglou[*], Ph.D.; Olga Merport[*], M.S.;
Kathleen R. McKeown[*], Ph.D.; and **Desmond A. Jordan[†], M.D.**

[*]Department of Computer Science
Columbia University, New York, NY 10027

[†]Departments of Anesthesiology and Medical Informatics
College of Physicians and Surgeons
Columbia University, New York, NY 10032

We present a representation model for the content of medical documents (journal articles and the patient's record) that allows the extraction of critical relationship information from online texts and tabular data. Our model relies on a list of attributes with associated values that are dynamically determined using an efficient finite-state grammar and an automatic term verifier. Extracted relationships are partitioned in different subsets according to the particular group of patients they refer to, thus enabling the retrieval of multi-topic documents and the targeted selection and presentation of a portion of a document's information. We present results from a system implementing this representation model and contrast our approach with traditional information retrieval and information extraction techniques.

INTRODUCTION

The rapid expansion of online information in the past few years has brought about a corresponding increase in the amount of medical content available through the Internet. A number of medical journals (e.g., the American Heart Journal) make articles available electronically, either on a subscriber basis or to the public at large; government, professional associations, and private concerns maintain online repositories (of varying quality) of medical information targeted to health care personnel. In addition, hospitals frequently maintain their own electronic databases of patient information, accessible over the network to health care providers practicing there.

This wealth of online information has meant that physicians need to spend an increasing amount of time browsing online information in order to keep informed of developments in their field. In addition, physicians may use online searches to seek specific information applying to a particular patient under their care. This is frequently the only viable means for locating information outside a person's speciality; for example, a cardiac anesthesiologist may regularly follow the five journals specializing in this area, but cannot be expected to track papers from all of the 60 journals in cardiology, the 40 journals in cardiothoracic surgery, and, even less, the more than 1,000 journals in the broader field of internal medicine. It is not uncommon for physicians in training (residents and interns) to spend several hours per day browsing online sources.

Yet, current information retrieval and presentation technology does not provide the level of targeted information selection that would be most helpful during these searches. Responses to queries in a search engine such as MEDLINE [10] contain extraneous articles that are not truly relevant (false positives), and miss several of the articles that are relevant. More importantly, the granularity of the responses is too coarse: At best, the physician is presented with several potentially relevant articles, and has to read through them to find the parts that apply to his or her current patient. The presentation of results at the article level makes the collation and summarization of related information (for example, experimental results on comparable patients) hard.

In this paper, we present a new model for representing the information from an online source, be it an article in a medical journal or a laboratory report in an online clinical information system. Rather than abstracting the content of an online document through a list of frequent keywords, as traditional search engines and text matching algorithms do, we represent content as a list of patient-specific, medically informed *attributes* with associated *values*. For example, an article may broadly refer to two patient populations, sharing some common characteristics and distinguished on others; these characteristics can represent demographics, pre-existing medical conditions, laboratory results, treatments, etc. We elaborate on the concept of attribute-value lists and argue for their superiority as a representation model in the next section, contrasting with the word vector model used in related work. Then, we present a fully implemented system for constructing these lists out of textual and tabular data, and show how the lists can be partitioned to correspond to multiple patient populations discussed in the same document. Results from our system are presented, and we conclude by discussing a number of related issues that are the focus of our current research.

THE ATTRIBUTE-VALUE REPRESENTATION

Traditional information retrieval (IR) engines index documents by representing them as a vector of word

frequencies modified by the overall rarity of each word in the domain or the language at large (rare words count more towards a match; very common words are completely filtered out using a stoplist). This TF*IDF (total frequency/inverse document frequency) approach [13, 3] offers a number of general advantages: it automatically focuses on the most important words in the document; it can be applied in any domain and can be retrained on a new collection from a specific domain to automatically adapt the rarity (IDF) factors; and it requires minimal analysis of the documents, typically limited to word stemming and normalization of capitalization. Yet, when applied to the medical domain, the IR approach suffers from two major drawbacks:

- Because the indexing is done on a word-by-word basis, common words that are significant in a particular context are ignored. This has been recognized as a problem in the information retrieval community, and solutions have been proposed to capture pairs (or longer sequences) of *fixed* words such as technical terms [14]. In this manner, "heart failure" will be assigned a higher weight as a single entity rather than relying on matching the individual (rather frequent) words "heart" and "failure". While this extension to the word vector approach may be adequate for such collocations and technical terms, it still fails to capture a large number of word combinations where any member of an open set of expressions (e.g., a number) is employed. For example, the phrase "blood pressure less than 90" could be crucial in selecting or rejecting a document if it specifies a mandatory inclusion or exclusion criterion for a study; but information retrieval systems will, at best, only recognize and use "blood pressure" from the above expression.
- Since matching and retrieval is performed at the document level, a document that contains a part relevant to a query but is mostly about something else is likely to be omitted from the results. This is a particular problem if a document refers to multiple topics (e.g., multiple disjoint patient populations with different defining characteristics). Not only is a relevant portion of a document likely to be ignored if it does not represent a significant part of the overall document, but the output of IR systems also presents the retrieved documents in their entirety. Thus, the user runs the risk of missing a multitopic document, and even when he or she is presented with one, has to shift through the entire text to locate the part that is really relevant to the original query.

Given these significant deficiencies of current information retrieval technology, an alternative approach has been pursued, based on the *semantic analysis* of the input documents. Such *information extraction* systems perform a deeper analysis of the source text, using some form of parsing to recover linguistic structure and a preconstructed model of the semantics in the domain. Often, the parsing and semantic analysis are combined in a semantic grammar that employs semantic categories (such as "laboratory test result") as nonterminals. The systems developed for the Message Understanding Conferences (MUC) [15] under the auspices of DARPA[1] are a good example of a sustained, multi-year, and multi-site effort in this arena, dealing with information extraction from news articles reporting terrorist events; similar systems for medical subdomains (e.g., [1, 12, 4, 5]) have also been developed.

Systems of this type are able to locate expressions such as the blood pressure example above and understand their meaning sufficiently well to consider them important for retrieval; they can also support (at least in principle) targeted retrieval of smaller portions of an input document. However, they suffer from major drawbacks of scalability and portability: Because they require a semantic model of the domain they operate in, they are limited in the type of semantic categories they cover, they take a long time and significant amount of domain expertise to develop, and are not portable to other domains without major revisions. The MUC systems mentioned above, for example, can analyze a news story about a bombing but not one about retaliatory actions and countermeasures taken against terrorists. Similarly, medical information extraction systems need to be limited to a narrow domain (e.g., radiology reports) to be effective.

We propose a new model for representing the content of a document that combines the benefits of the information retrieval and information extraction approaches without their disadvantages. Our model represents information in the form of attribute-value pairs, where *attribute* is a (possibly multi-word) expression describing a medical term (disease, symptom, treatment, laboratory test, etc.) and *value* is an (again possibly multi-word) expression describing the outcome associated for that term. Depending on the attribute, the value may be implicit (i.e., presence or absence of a disease, such as in the term "diabetic"), a single numeric or text value (e.g., value "III" for the attribute "New York Heart Association Class"), a fixed or open-ended range of values (e.g., "aged 30 to 60 years", "blood pressure less than 90"), or a qualitative description (e.g., value "severe" for the attribute "heart failure").[2] Like the information extraction systems described earlier, our model captures important relationships between words and terms that IR systems cannot capture. But unlike the information extraction systems, our approach does not involve building a domain-specific semantic model; instead, valid attributes are collected on the fly using a combination of general syntactic patterns and

[1] The Defense Department's Advanced Research Projects Agency

[2] All the above examples are actually extracted by our system from the collection of articles described in the results section later.

a term verification component. Hence we achieve generality and portability without sacrificing the ability to recognize important relationships.

In order to handle multi-topic documents, we partition the attribute-value pairs extracted from a document according to the group of patients they apply to. We recognize expressions denoting such groups and identify their scope. Text within each group's scope is analyzed for valid attribute-value pairs. Pairs from all such group scopes are clustered together by matching group descriptions that refer to the same patient population (e.g., "the control group"). The result of our system is a set of attribute-value lists, one for each distinct group mentioned in the source document.

EXTRACTING ATTRIBUTE-VALUE PAIRS

For clarity, we discuss first the simple case where the source document contains information about a single patient or group of patients. This is the case, for example, in laboratory tests and pre- and post-operative reports maintained as part of a patient's record in an online clinical information database. When actually processing a document from a medical journal, which could refer to multiple patient populations, our system first applies the group identification and stratification procedure of the next section, then follows the algorithm below for the parts of the document that fall within each group's scope.

We first recognize and process tabular data, which can be signalled by certain types of indentation in ASCII sources and by appropriate HTML tags in documents stored on the Web. Such data is common in the various patient reports, and occasionally occurs within articles as well. Information from the row and column names is used to construct composite attribute names which are assigned the corresponding cell's contents as their value.

The remainder of the source text that does not match our specifications for a table is treated as text. Our algorithm for collecting attributes and their associated values relies on finite state automata to recognize syntactic constructs indicating relationships between (potential) medical terms and optional values. We start by assigning part-of-speech information using a publicly available tagger [2]. Then we recognize simple noun phrases, consisting of an optional sequence of adjectives followed by one or more nouns. Patterns of simple noun phrase combinations, as well as combinations of simple noun phrases and numbers, relative expressions, and certain prepositions are then extracted, according to the rules listed in Table 1. In addition to the rules given in Table 1, we also recognize passive forms and conjunctions of adjectives which are transformed into multiple simpler forms.

This approach recovers a large number of potential attributes and their values, but of course overgenerates by collecting many non-medical terms that match the definition of a simple noun phrase. Furthermore, sometimes the attribute and its value are embedded within the same simple noun phrase (e.g., "severe heart failure"). To address both of these concerns, we check each of the candidate attributes with a *term verifier* and keep only those that are recognized as terms in the medical domain. Our term verifier relies on the Unified Medical Language System (UMLS) [8], a front-end to several large-scale medical knowledge bases. Terms that are found in the database are considered valid attributes; in addition, our verifier supports partial matching of a term in the database as a right substring of a candidate attribute. In that case, the non-matched portion at the start of the attribute becomes the value, thus properly separating "severe" from "heart failure". Note that by changing the referenced knowledge base, which is quite broad to start with, we can easily port our system to a different domain.

EXTRACTING GROUP INFORMATION

Our system stratifies attribute-value pairs according to the groups of patients they apply to. We start by recognizing group expressions, using again finite state automata on text that has been assigned part of speech information. In addition to general rules involving broad syntactic categories, the group recognition automata also utilize specific key words (such as "group" or "patients"). This allows for a much greater variety of rules, but carries the risk of over-adapting the group recognition to the limited set of articles used for system development. We address this problem by adopting a semi-automatic approach where group patterns are automatically learned from the articles: We start with a few *seed* words (such as the examples given above, "group" and "patients"), and automatically extract all word sequences in windows of width up to five on each side of a seed word; these expressions are then collated and manually filtered to select general group expressions. The selected expressions usually suggest other promising seed words, so the process is repeated until no more general expressions can be found. Figure 1 shows some of the rules extracted in this manner from a set of journal articles in cardiology.

Our grammar for group recognition partitions a group specifier into an essential part and an (optional) additional qualifier; for example, in the group specifier "40% of control group patients", the essential part is "control group patients" while the qualifier "40%" does not play a role in distinguishing this group from other groups mentioned in the same document. We collate information from different occurrences of the same underlying group by matching the essential parts of the group specifiers.

Finally, we determine the scope of each group specifier by separating successive clauses using punctuation and *wh*-words introducing relative clauses. We recognize a

Pattern	Example
NP with at least one adjective	*Severe conjestive heart failure*
NP + comparative-operator + (AP or NP or Q)	*Base heart rate less than 90*
NP + of + [comparative-operator] + Number	*Blood pressure of (more than) 80*
NP + comparative-adjective + than + Number	*Left ejection fraction lower than 20%*
NP + linking-verb + (AP or NP or Q)	*The 2-year actuarial mortality rate was 48%*

Table 1: Rules recognizing potential attributes and their associated values. NP stands for a simple noun phrase, AP for one or more adjectives, and Q for a quantitative expression involving a number and possibly additional nouns, numbers, and prepositions.

<People> whose <NP> was <Value>
<People> with <NP>
(e.g., <NP> = *base heart late less than 90)*
<People> after <NP>
(e.g., <NP> = *myocardial infraction)*
<People> undergoing <NP>
(e.g., <NP> = *coronary angioplasty)*
<People> older than <Number> years

Figure 1: Some of the group recognition patterns extracted semi-automatically from cardiology medical articles.

Attribute	Value
amiodarone	effective
congestive heart failure	severe
left ventricular ejection fraction	35%
mortality rate	48%
ventricular tachycardia	asymptomatic

Table 2: Partial list of attribute-value pairs for one of the groups discussed in an article on amiodarone treatment of congestive heart failure.

number of composite group patterns, such as comparative sentences of the form "GROUP1=<Men> [were receiving antiplatelet therapy] CONTRAST=<more frequently than> GROUP2=<women>". By default, the scope of a group extends until a new group specifier is recognized (thus allowing a group's scope to span multiple sentences).

RESULTS

We have tested our system on a set of 185 medical articles, collected from two online journals ("Journal of the American College of Cardiology" and "American Heart Journal"), as well as on a number of online patient records available through the clinical information system at Columbia-Presbyterian Medical Center [11]. We show partial representative results from the attribute-value extraction process for one group discussed in a sample cardiology article [9] (Table 2); Figure 2 shows some of the group expressions and the text associated with them from the same article.

We are currently working on planning a formal evaluation of these results. We intend to measure quantitatively the accuracy of components of our system (e.g., percentage of overall attribute-value pairs that are recovered, rate of false positives, accuracy of our term verifier) as well as empirically compare our representation and system to standard search engines in a task-based evaluation.

The mortality rate was
CONTRAST: lower
in
GROUP1: amiodarone-treated patients
COMPARATIVE: compared with
GROUP2: control patients.

A total of
GROUP1: [QUALIFIER: 367] patients
had their rest heart rate determined at 6 months of follow-up.

Figure 2: Some of the group specifiers and matched text recognized from one medical article on amiodarone treatment of congestive heart failure.

CONCLUSION AND FUTURE WORK

The task of finding a concise, yet adequate, representation for the content of medical documents is non-trivial and gives rise to a number of interesting side issues. We have proposed a novel representation in terms of flexible attribute-value pairs, which, although related to the slot-filler representation used in information extraction systems, differs from them in that it contains no prespecified semantic component. We have discussed techniques for extracting this type of information in a general way that can be easily ported to different domains, and presented results that indicate the viability of this approach for intelligent matching of documents to queries, other documents, or a patient's online record.

Our current work involves improvements in some of the components of the system, as well as addressing related issues that would enhance its usefulness. A major area of our work is the problem of matching two attribute-value lists, so that a document can be matched to a particular patient, for example. This task is complicated because not all attributes should carry

the same weight (demographic data could be less significant), and even the matching of values for a single attribute is non-trivial when qualitative descriptions such as "severe" are used. We are also interested in improving our term verifier—while the UMLS provides an extensive list of terms, it is unfortunately incomplete and contains non-medical noun phrases (such as "African American"). We are investigating techniques that combine statistical term verification [6] with rules that employ the semantic classes provided by UMLS. Finally, our original motivation for this system has been to facilitate multi-document summarization that goes beyond simple sentence extraction [7]. This last task is supported by our extraction of information targeted to specific patient groups, and we are investigating ways to combine the matching text fragments into a concise and fluent summary that identifies repetitions and contradictions across the multiple source documents.

Acknowledgements

We thank CPMC cardiac anesthesia physicians Shabina Ahmad, Terry Koch, Jonathan Oster, and Carmen Domingez for early input during the design of this system. Jonathan Oster has in addition marked source documents with the information that our system should aim to extract. Kathleen Dunn, of the Department of Computer Science, implemented our local installation of the UMLS database as a Sybase server. Federico Kattan helped with the collection of medical articles and abstracts. The research work is supported in part by the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare, funded by the New York State Science and Technology Foundation. Any opinions, findings, or recommendations are those of the authors, and do not necessarily reflect the views of NYSSTF.

References

1. R. Baud, A. Rassinoux, and Jean R. Scherrer. Natural language processing and semantical representation of medical texts. *Methods of Information in Medicine*, **31**(2):117–125, 1992.
2. Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992. Association for Computational Linguistics.
3. William B. Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, New Jersey, 1992.
4. Carol Friedman, George Hripcsak, William DuMouchel, Stephen B. Johnson, and Paul D. Clayton. Natural language processing in an operational clinical information system. *Natural Language Engineering*, **1**(1):83–108, 1995.
5. George Hripcsak, G. J. Juperman, and Carol Friedman. Extracting findings from narrative reports: Software transferability and sources of physician disagreement. *Methods of Information in Medicine*, **37**:1–7, 1998.
6. John S. Justeson and Slava M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, **1**(1):9–27, 1995.
7. Kathleen R. McKeown, Desmond A. Jordan, and Vasileios Hatzivassiloglou. Generating patient-specific summaries of online literature. In *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Text Summarization*, pages 34–43, Stanford, California, March 1998. American Association for Artificial Intelligence.
8. National Library of Medicine, Bethesda, Maryland. *Unified Medical Language System (UMLS) Knowledge Sources*, sixth experimental edition, 1995. Accessible at `http://umlsks.nlm.nih.gov`.
9. Daniel R. Nul, Hernan C. Doval, Hugo O. Grancelli, Sergio D. Varini, Saul Soifer, Sergio V. Perrone, Noemi Prieto, and Omar Scapin. Heart rate is a marker of amiodarone mortality reduction in severe heart failure. *Journal of the American College of Cardiology*, **29**(6):1199–205, May 1997.
10. National Library of Medicine. MEDLINE. World Wide Web site, URL: http://www.nlm.nih.gov/databases/medline.html. Updated weekly January through October and twice in the month of December, 1997.
11. N. Roderer and P. Clayton. IAIMS at Columbia Presbyterian Medical Center: Accomplishments and challenges. *Bulletin of the American Medical Libraries Association*, pages 253–262, July 1992.
12. Naomi Sager, Margaret S. Lyman, C. Bucknall, N. T. Nhan, and L. J. Tick. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Society*, **1**(2):142–160, 1994.
13. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **25**(5):513–523, 1988.
14. Alan F. Smeaton. Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, **35**(3):268–278, 1992.
15. Beth M. Sundheim. Overview of the fourth message understanding evaluation and conference. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)*, pages 3–21, McLean, Virginia, June 1992. DARPA Software and Intelligent Systems Technology Office, Morgan Kaufmann, San Mateo, California.