

Converting Semi-structured Clinical Medical Records into Information and Knowledge

Xiaohua Zhou¹, Hyoil Han¹, Isaac Chankai², Ann A. Prestrud² and Ari D. Brooks²

College of Information Science and Technology, Drexel University¹

Department of Surgery, College of Medicine, Drexel University²

3141 Chestnut Street, Philadelphia, PA 19104, USA

{Xiaohua.Zhou, hyoil.han}@drexel.edu, {ic36, Ann.Prestrud, Ari.Brooks}@DrexelMed.edu

Abstract

Clinical medical records contain a wealth of information, largely in free-textual form. Thus, means to extract structured information from free-text records becomes an important research endeavor. In this paper, we propose and implement an information extraction system that extracts three types of information — numeric values, medical terms and categorical value — from semi-structured patient records. Three approaches are proposed to solve the problems posed by each of the three types of values, respectively, and very good performance (precision and recall) is achieved. A novel link-grammar based approach was invented to associate feature and number in a sentence, and extremely high accuracy was achieved. A simple but efficient approach, using POS-based pattern and domain ontology, was adopted to extract medical terms of interest. Finally, an NLP-based feature extraction method coupled with an ID3-based decision tree is used to classify and extract categorical cases. This preliminary approach to categorical fields has, so far, proven to be quite effective.

1. Introduction

Patient medical records contain a wealth of information that can prove invaluable for the conduct of clinical research. Clinical records are largely maintained in free-text form. Thus, a reliable, efficient method to extract structured information for future

data mining from free-texts using information extraction techniques may greatly benefit research endeavors.

Management of medical information is an essential aspect of good patient care. Most clinicians have standardized systems for documenting patient visits and procedures, almost entirely paper-based. Some clinicians "take notes"; directly handwriting information in a patient's chart. Others dictate information, which is subsequently typewritten and may or may not also write notes in charts. Most terminology and abbreviations are generally uniform and can be easily understood by all doctors; meaning a cancer surgeon can easily communicate, through patient notes or dictated letters, with a cardiologist. However, many abbreviations for singular terms exist. Manual processing of patient information into a database is subject to errors and may not always maintain objectivity, as well as being expensive. These are some reasons why electronic medical records have yet to be widely adopted in medical practice.

The medical community is constantly striving for new means to conduct research in the battle against diseases. One avenue frequently explored is chart review. This means of conducting a study is often fruitful, yet requires great attention to detail and is infinitely time-consuming. As a result, studies based on chart review are often limited, including a small number of cases. Means to systematically examine patient charts will provide a method for clinicians to examine a significantly larger set of cases. The value

of considering more records simultaneously is the ability to then detect small variations, which may pinpoint important factors previously overlooked. Information scientists have the tools and capability to provide such a method to expand the research lens. We report on development of information extraction and mining techniques that accurately identify desirable information from transcribed consultation notes. A total of 50 separate initial consultation notes are mined by the program. Results are then compared to a medical student's independent manual processing of the same 50 consultation notes.

In this paper, we propose and implement a system that can extract structured information from semi-structured patient records. The system reported herein is a part of a large research project on breast cancer being conducted in Drexel University's College of Medicine. Before researchers can conduct any analysis or mining, it is required that they code the textual patient records and save structured information into the database.

Information of interest in our project can be roughly classified into numeric values (e.g., blood pressure and pulse), medical terms (e.g., past medical history and past surgical history), and categorical values (e.g., a patient's status as a former or current smoker, or a nonsmoker).

Three approaches are proposed to extract these three types of values respectively. A novel link-grammar based approach was invented to extract numeric values. A simple but efficient approach, using POS-based pattern and domain ontology, was adopted to extract medical terms of interest. Finally, an NLP-based feature extraction method coupled with ID3-based decision tree is used to classify and extract categorical cases.

The remainder of this paper is organized as follows: in section 2, we review related work; in section 3 we present our own approaches to the extraction of the three types of information; section 4 describes the details of implementation; and section 5 evaluates the performance of the system. A short conclusion finishes the article.

2. Related Work

One line of research related to ours is Named Entity Recognition (NER) in free-text. Though most NER methods cannot handle medical terms directly, their idea, pattern matching, for example, can be borrowed. General Architecture for Text Engineering (GATE) [1] uses patterns written in regular expressions to implement all its components such as tokenization and

named entity recognition. It also provides a Java Annotated Pattern Engine (JAPE) [2], by which users can extend NER component to identify entities of interest. However, because medical terms are full of synonyms and morphologic variants, ontology is necessary to achieve high extraction accuracy for medical terms taken from clinical records. A research project, "Acquiring Medical and Biological Information from Text" (AMBIT) [3], led by a research group at the University of Sheffield, aims to build just such a large medical term database for the sake of information extraction from clinical records. In this particular project, we employ Unified Medical Language System (UMLS)¹ as the domain ontology to identify medical terms.

Linguistic pattern-based template filling is a common technique for information extraction. AutoSlog [9], PALKA [4], CRYSTAL [13] and WHISK [14] all can automatically induce linguistic patterns from training examples. However, supervised pattern learning is costly. Instead, we use an unsupervised approach, which makes use of the results of link grammar parser [11], to extract a portion of knowledge in the project.

Another line of related research is text classification. Decision trees are a frequently used technique for text classification. Wendy Lehnert et al. [6] present an ID3-based decision tree for classification, which uses learned keywords as features [6]. Roland Kuhn and Renato De Mori propose application of semantic classification trees (SCT) to natural language understanding [5]. SCT is an extension to word-based (as feature) decision trees. Unlike [6] and [5], Riloff and Lehnert [10] describe an approach to text classification that represents a compromise between word-based technique and in-depth natural language processing. It takes polysemy, synonyms, phrases, and local context into account during feature extraction.

3. Methods

Information of interest in our project can be roughly classified into numeric values (e.g., blood pressure and pulse), medical terms (e.g., past medical history and past surgical history), and categorical values (e.g., a patient's status as a former or current smoker, or a nonsmoker). For different types of information, it is difficult to use the same method to achieve good performance. Instead, we apply an analytic approach (NLP technique plus domain ontology) to the

¹ <http://www.nlm.nih.gov/research/umls/>

extraction of the first two types of information, and we approach categorical values with NLP techniques plus supervised machine learning. Appendix shows typical examples of clinical medical records that were used in this project.

3.1 Approaches to Numeric Fields. One type of information of interest is number. For example, a patient's blood pressure, pulse, age and weight. In our project, subjects are patients with breast cancer; numeric fields of interest also include menarche age, number of pregnancies, number of live births, etc.

Numbers in patient records can be either digits (e.g. 17) or English words (e.g., seventeen), either of which we can easily extract. In fact, most NLP development tools, such as GATE, provide tokenization modules and Named Entity Recognition modules, which annotate all numbers in a text with extremely high precision and recall.

The identification of a feature (field or attribute) name in a text is not difficult either. One straightforward approach is an exact text search of the feature name. In order to improve the recall of feature identification, we further introduce target synonyms and inflected variants of the feature and its synonyms. Currently, we are manually specifying the synonyms of the concept. In the future, we will automate this part as synonym databases of biomedical terms are publicly available online. Regarding inflected variants, we used WordNet and some heuristics to automatically generate them from original concepts.

However, it is difficult to associate the extracted number with identified features because in the majority of cases a sentence contains more than one feature. For example, the following sentence contains four features: blood pressure, pulse, temperature and weight.

Blood pressure is 144/90, pulse of 84, temperature of 98.3, and weight of 154 pounds.

One shallow approach to the association between numbers and features is through the use of linguistic patterns or heuristics. Some examples of linguistic patterns are listed below:

- (1) **CONCEPT is NUMBER**
- (2) **CONCEPT of NUMBER**
- (3) **CONCEPT, NUMBER**
- (4) **CONCEPT: NUMBER**

The major advantage of a pattern approach is its simplicity. However, this approach has generalization problems because the expression of natural language is so flexible. Here we propose and implement a novel approach, which uses the linkage information

produced by Link Grammar Parser [11] to associate between numbers and features.

Link Grammar is an original sentence parser, producing not only a constituent tree as most parsers yield, but also a linkage diagram that consists of links between two words. In Figure 1, there are 4 links. The link between “is” and “144/90” represents a verb-object relation (denoted by notation ‘O’). Suppose a node represents a word, and an edge represents a link. Then the linkage diagram of a valid sentence can be looked at as a connected graph. Furthermore, each edge can be weighted against the type of link according to the application. Thus, the shortest distance between any word pair can be calculated from the graph.

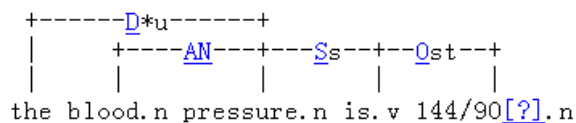


Figure 1. An Example of a Linkage Diagram²

Intuitively, the shortest distance between any word pair is a good measure of the semantic relationship of the word pair. Specifically, we assume that a feature's numeric value will be the number that exists closest to the feature keyword in the sentence. Therefore, the association of feature and number in a sentence is equivalent to searching for the node (feature) with the shortest distance from a fixed node (number) in a (weighted) graph.

One limitation of this analytic approach is that the Link Grammar Parser cannot parse text fragments, e.g., “blood pressure: 144/90.” Thus, we also implement the pattern approach. If the parser fails to parse the sentence, the pattern approach is used.

3.2 Approaches to Medical Terms. Another type of information of interest is medical terms such as past medical history, past surgical history, and symptoms of a certain disease. Many medical terms are multi-word phrases. For the following description of past medical history, the system will extract three terms: *postoperative CVA*, *cholecystectomy*, and *midline hernia*.

"Significant for a postoperative CVA after undergoing a cholecystectomy and a midline hernia closure"

² The diagram is yielded by an online Link Grammar parser at <http://www.link.cs.cmu.edu/link/>.

Because medical terms are often multi-word phrases, it is not efficient to search all combinations of sequential words in the sentence from certain domain ontology (medical vocabulary). Instead, we use part of speech tagger [1] and ordered patterns to obtain a list of term candidates, and then see if the candidate terms exist in the ontology. The approach is illustrated by the example of extracting past medical history.

The first step is to tag the part of speech for the words in each sentence. Then we employ the following four ordered patterns to find candidate terms Here JJ and NN denote adjective and noun respectively. The first pattern, for example, matches a three-word term that is comprised of an adjective and two nouns in order.

- (1) JJ NN NN
- (2) NN NN
- (3) JJ NN
- (4) NN

Finally, we search through UMLS (a medical vocabulary database). If a term exists in the database, we then save it and continue to look for terms after the current term's endpoint. Otherwise, we look for terms matching the next pattern from the current starting point.

Moreover, before searching through the database, it is often necessary to normalize the candidate term because a term has a variety of surface forms in text but is often stored in unique normalized form in database for the purpose of efficient query. Normalization usually includes two steps: (1) getting the uninfected form of the surface word, (2) sorting multiple words in alphabetic order. For example, the term "high blood pressures" after normalization becomes "blood high pressure." The normalization can be easily implemented by using WordNet [7].

3.3 Approaches to Numeric Fields. The primary type of information in our project is categorical value. For instance, smoking behavior has three categorical values: never, former, and current. The shape of a patient is classified into four categories: thin, normal, overweight, and obese. The following texts are examples of patient smoking behavior.

"She quit smoking five years ago" (former)

"She is currently a smoker" (current)

"None" (never)

"She has never smoked" (never)

For high accuracy, an analytic NLP approach is recommended by most available literature. Usually, pattern-based semantic analysis would be performed to classify the cases. However, the analytic approach

highly demands large amounts of domain knowledge, and is consequently difficult to generalize.

Conversely, a machine learning technique does not depend on domain knowledge, and the approach can easily be generalized. In this project, we employ an ID3-based decision tree [8] for categorical fields. According to information theory, Information Gain (Mutual Information) of the predictor and dependent variable is a good measure of the predictor's discriminating ability. Thus, the ID3 decision tree is supposed to use less features than other decision tree algorithms.

In order to achieve high accuracy of classification, it is crucial to extract informative candidate features. In the field of NLP, features are usually the presence or absence of a certain word or phrase. In this project, the presence of a certain word is treated as a Boolean feature. To lessen the computing burden and increase the use of domain knowledge, our method for feature extraction allows the following options to be chosen by the user for each field for extraction.

- (1) Choose one or multiple part of speeches: verb, noun, adjective, and adverb.
- (2) Choose one or multiple sentence constituents: subject, verb, object, and supplement.
- (3) Head noun or head adjective only. If this option is enabled, for a noun phrase or an adjective phrase, only the head word is extracted.
- (4) Use lemma (uninfected form) of any word. If this option is enabled, "denies," "denied" and "deny" will be treated as the same feature. The use of lemma will not only reduce the number of candidate features, but also influence the choice of nodes during the construction of a decision tree. We recommend enabling this option unless domain knowledge suggests that the infected form is indicative of classifications.

In smoking behavior classification, for example, we search for certain parts of speech — verbs, nouns, adjectives, or adverbs — that appear in any constituent part of the sentence; meanwhile, we disable the "head noun or head adjective only" option, and enable the "use of lemma" option. We chose the above parameter settings out of our understanding to the particular task and also in part the experiment results.

The above feature extraction method works well for most categorical fields in our project. But for classifications containing numeric information, performance is poor. For example, alcohol use has four classes: never, social, 1-2 day per week, >2 day per week. It is reasonable to gain poor performance because not all numbers in the range of interest will

appear in training cases. To solve this problem, we plan to add one more type of feature — a numeric Boolean feature — to the next version. In the example of alcohol use, two more added features are: (1) whether a number less than or equal to 2 appears in the sentence, (2) whether a number greater than 2 appears in the sentence. The thresholds are manually specified by users.

4. Implementation

The system is implemented by Java. For external resources such as Link Grammar parser and WordNet written in native C code, we access their functions via third-party Java Native Interface (JNI). The system architecture is shown in Figure 2.

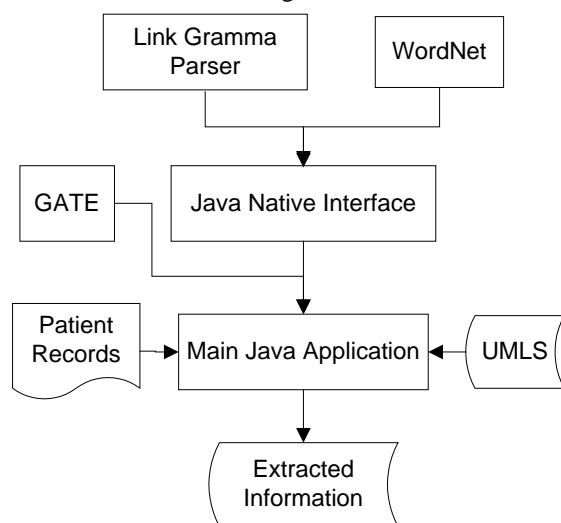


Figure 2. System Architecture

Link Grammar Parser³ is used to produce both linkage information for the association of number and feature and a constituent tree for feature extraction.

WordNet⁴ is mainly used to get the lemma (uninflected form) of each surface word in a sentence.

GATE⁵ (General Architecture for Text Engineering) is used for tokenization, sentence splitting, and part of speech tagging. After tokenization, all numbers in the text are identified.

³ The source code and installation package for Link Grammar parser version 4.1 are available at: <http://www.link.cs.cmu.edu/link/ftp.html>. Its Java Native Interface can be downloaded from the address: <http://www.chrisjordan.ca/research/LGInterface.tar>.

⁴ The installation package of WordNet 2.0 is available at <http://wordnet.princeton.edu/>. Its JNI can be downloaded at <http://wnjn.sourceforge.net/>.

⁵ <http://gate.ac.uk/>

UMLS serves as the domain ontology for searching medical terms. For the sake of efficiency, we downloaded UMLS data and installed it in a local DB2 database. The data is accessed by JDBC.

Patient records for input are stored in separate ASCII text files. Extracted information is saved in a Microsoft Access database. We implemented the ID3-based decision tree algorithm.

5. Evaluation

The data set consists of fifty patient records, each of which is for a subject with breast complaints. The data format is semi-structured and shown in the Appendix. One record is comprised of multiple sections, each of which begins with a fixed string. Therefore, it is easy to split the whole record into sections. Each section is written in natural language.

The task is to extract eighteen fields from the text. Some fields contain more than one attribute. The extraction of twenty-four attributes in total is required, among which are four numeric multi-valued medical terms, eight numeric attributes, and twelve categorical attributes. Among the twelve categorical attributes, six are binary classifications. So far, we completed all four medical term attributes, all eight numeric attributes, and one categorical attribute (smoking). The evaluation was performed on attributes completed.

We used two measures, precision and recall, to evaluate performance of the information extraction system. Precision is defined as the proportion of correctly extracted instances of those extracted, while recall is the proportion of correctly extracted instances of total instances.

Precision (recall) for all eight numeric attributes is 100%. By examining all fifty records manually, we find that the extremely high precision can in part be attributed to the very consistent dictation style (all records were provided by the same clinician, the author Ari D. Brooks, MD). If the size of the data set increases or the writing style is full of variants, performance may be degraded.

The ID3-based decision tree is evaluated on the attribute of smoking behavior. Because five subjects don't have smoking information, forty-five cases are evaluated; among which five are former smokers, twelve currently smoke, and twenty-eight never smoked. Five-fold cross validation is applied. That is, the whole data set is split into five subsets; for each round, four subsets are treated as training data and the last one as testing data. We run a five-fold cross validation ten times, and each time the dataset is randomly shuffled. Average precision (recall) is

92.2%. The number of features used in the decision tree ranges from four to seven.

Past medical history and past surgical history have multiple medical term values. Thus, the precision and recall for i -th subject are defined respectively as:

$$R_i = \frac{ETrue_i}{TInst_i} \quad P_i = \frac{ETrue_i}{ETotal_i}$$

Precision and recall for all cases, respectively, are defined as below:

$$R = \frac{\sum_i ETrue_i}{\sum_i TInst_i} \quad P = \frac{\sum_i ETrue_i}{\sum_i ETotal_i}$$

Where:

$ETrue_i$: number of extracted true terms in i -th subject.

$ETotal_i$: number of extracted terms in i -th subject.

$TInst_i$: number of total true terms in i -th subject.

The precision and recall for the extraction of four multi-valued medical terms is listed in Table 1.

Table 1. The Performance of Medical Term Extraction

Attribute Name	Precision	Recall
Predefined Past Medical History	96.7%	96.7%
Other Past Medical History	76.1%	86.4%
Predefined Past Surgical History	77.8%	35%
Other Past Surgical History	62.0%	75%

Analyzing results manually, we find that false positives are mainly caused by the incompleteness of domain ontology. Higher performance can be achieved by choosing an appropriate medical database. We also found that the low recall of predefined past surgical history and low precision of other past surgical history is due to failures to recognize the synonyms of predefined surgical terms and improper assignments of them to other surgical terms. This problem can be solved by introducing synonyms.

6. Conclusions

In this paper, we proposed and implemented an information extraction system that extracts three types of information — numeric values, medical terms and categorical values — from semi-structured patient records. Three approaches are proposed to solve the problems posed by the three types of values,

respectively, and very good performance (precision and recall) is achieved.

A link-grammar based novel approach was invented to associate feature and number in a sentence and extremely high accuracy was achieved. A simple but efficient approach using POS-based pattern and domain ontology is adopted to extract medical terms of interest. Finally, an NLP-based feature extraction method coupled with an ID3-based decision tree is used to classify and extract categorical cases. Our preliminary analysis with this approach to categorical fields has proven to be quite good.

However, the size of the data set is small. When more diversified writing styles are introduced into patient records, the performance of the extraction process may be degraded. We will need to use a larger data set to reevaluate and tune our future work. Moreover, we have not completed classification of all categorical fields, particularly fields containing numeric information to which we propose an adaptive approach. In addition, for medical term fields, there is still a room to improve by choosing an appropriate ontology (medical database).

This approach may offer a new means by which clinicians may extract large volumes of data from patient medical records. To date, this resource is not tapped as there is no effective means to extract data. We hope to continue this work, refining our approach, to expand its utility.

References

- [1] Cunningham, H., "GATE, A General Architecture for Text Engineering", *Computers and the Humanities*, 2002, Vol. 36, pp. 223-254
- [2] Cunningham, H., Maynard, D., and Tablan., V., "JAPE: a Java Annotation Patterns Engine (Second Edition)", Technical report CS--00--10, University of Sheffield, Department of Computer Science, 2000.
- [3] Gaizauskas, R., Hepple, M., Davis, N., Guo, Y., Harkema, H, Roberts, A., and Roberts, I., "AMBIT: Acquiring Medical and Biological Information from Text", *ISMB/ECCB, Poster*, 2004.
- [4] Kim, J.T. and Moldovan, D.I., "Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction", *IEEE Transactions on Knowledge and Data Engineering*, Volume 7, Issue 5, 1995, pp. 713-724.
- [5] Kuhn, R. and Mori, R., "Application of Semantic Classification Trees to Natural Language Understanding", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, Vol. 17, No. 5.

- [6] Lehnert, W., Soderland, S., Aronow, D., Feng, F., and Shmueli, A., "Inductive Text Classification for Medical Applications", *Journal for Experimental and Theoretical Artificial Intelligence*, 1994, 7(1), pp. 49-80.
- [7] Miller, G. et al, "WordNet: an On-line Lexical Database", *International Journal of Lexicography*, 1990, pp. 235-245.
- [8] Quinlan, J.R., "Induction of Decision Trees", *Machine Learning*, 1986, No.1, pp.81-106.
- [9] Riloff, E., "Automatically Constructing a Dictionary for Information Extraction Tasks", *Proceedings of the Eleventh National Conference on Artificial Intelligence*, AAAI Press/the MIT Press, 1993, pp. 811-816
- [10] Riloff, E. and Lehnert, W., "Information Extraction as a Basis for High-Precision Text Classification ", *ACM Transactions on Information Systems*, 1994, Vol. 12, No. 3, pp. 296 – 333.
- [11] Sleator, D. and Temperley D., "Parsing English with a Link Grammar", *Third International Workshop on Parsing Technologies*, 1993.
- [12] Soderland, S., Aronow, D., Fisher, D., Aseltine, J., and Lehnert, W., "Machine Learning of Text Analysis Rules for Clinical Records", CIIR Technical Report, University of Massachusetts Amherst, 1995.
- [13] Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W., "CRYSTAL: Inducing a Conceptual Dictionary", *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1995, pp. 1314-1319.
- [14] Soderland, S., "Learning Information Extraction rules for Semi-structured and free text", *Machine Learning*, Vol. 34, 1998, pp. 233-272.

Appendix

The following shows a typical example of clinical medical records that were used in the project.

Patient: 2

Chief Complaint: Abnormal mammogram.

History of Present Illness: Ms. 2 is a 50-year-old woman who underwent a screening mammogram, revealing a solid lesion as well as an abnormal calcification. This was evaluated with further views including an ultrasound and a BIRAD 4. Classification was given. She was referred for further management. Her breast history is negative for any previous biopsies or masses.

GYN History: Menarche at age 10, gravida 4, para 3, last menstrual period about a year ago. First live birth at age 18.

Past Medical History: Significant for diabetes, heart disease, high blood pressure, hypercholesterolemia, bronchitis, arrhythmia, and depression.

Past Surgical History: Cervical laminectomy.

Medications: Aspirin, hydrochlorothiazide, Lipitor, Cardizem, senna, Wellbutrin, Zoloft, Protonix, Glucophage, Os-Cal, Combivent, and Flovent.

Allergies: Penicillin, ACE inhibitors, and latex.

Social History: Smoking history, 15 years. Alcohol use, occasional. Drug use, significant for marijuana.

Family History: Mother with breast cancer, diagnosed at age 52. Maternal aunt with breast cancer. No other family members with cancers.

Review of Systems: Significant for back pain and arthritis complaints. Also, allergies as listed above. Breathing issues are related to COPD, smoking, and diabetes. Remainder of the review of systems is negative.

Physical examination: Reveals an overweight woman in no apparent distress.

Vitals: Blood pressure is 142/78, pulse of 96, and weight of 211.

HEENT: PERRLA.

Neck: There is no cervical or supraclavicular lymphadenopathy.

Chest: Clear to auscultation anteriorly, posteriorly, and bilaterally.

Heart: S1 S2, regular, and no murmurs.

Abdomen: Soft, nontender, and no masses.

Examination of Breasts: Shows good symmetry bilaterally. Palpation of both breasts shows no dominant lesions. There is no axillary adenopathy.