

# Approaches for Information Extraction from Electronic Patient Records

Xiaohua Zhou

*College of Information Science and Technology, Drexel University  
3141 Chestnut Street, Philadelphia, PA 19104, USA  
Xiaohua.Zhou@drexel.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 implement an information extraction (IE) system that extracts a variety of information of patients with breast complaints from their semi-structured patient records. Three approaches are proposed to solve different IE tasks and very good performance (precision and recall) is achieved. A novel graph-based approach, which uses the parsing result of link-grammar parser, was invented for concept association, and extremely high accuracy was achieved. A simple but efficient ontology-based approach 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 perform text classification. 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.

Three approaches are proposed to solve different IE tasks and very good performance (precision and recall) is achieved. A novel graph-based approach, which uses the parsing result of link-grammar parser, was invented for concept association, and extremely high accuracy was achieved. A simple but efficient ontology-based approach 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 perform text classification. This preliminary approach to categorical fields has, so far, proven to be quite effective.

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) [5], 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 adopt Unified Medical Language System (UMLS)<sup>1</sup> as the domain ontology to identify medical terms.

Pattern-based template filling is a common technique for information extraction. AutoSlog [12], PALKA [6], CRYSTAL [16] and WHISK [17] can automatically induce linguistic patterns from training examples. However, supervised pattern learning is very expensive. Instead, we use an unsupervised approach, which makes use of the parsing results of

link grammar parser [14], to extract a good portion of knowledge in the project.

There is a bunch of research work that applies link grammar parser to information extraction. Madhyastha, Balakrishnan and Ramakrishnan reported the use of link grammar parser for event extraction [9] and Ding, Berleant, Xu and Fulmer applied link grammar to the extraction of biomedical interactions [4]. Both of their work reached the goal of information extraction by analyzing the meaning of important links in the sentence. Differing from their approaches, we first transform the parsed sentence to a formalism of graph and then perform concept association based on the generated graph.

Another line of related research is text classification. Decision trees are a frequently used technique for text classification. Wendy Lehnert et al. [8] present an ID3-based decision tree for classification, which uses learned keywords as features [8]. Roland Kuhn and Renato De Mori propose application of semantic classification trees (SCT) to natural language understanding [7]. SCT is an extension to word-based (as feature) decision trees. Unlike [8] and [7], Riloff and Lehnert [13] 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. Tasks and Methods

The information extraction tasks in our project can be roughly classified into three classes. The first one is the extraction of medical terms (e.g., past medical history and past surgical history). The second is text classification. For example, a patient falls into a former smoker, a current smoker, or a non-smoker). The last, also the major one, is about concept association (e.g. the association of symptoms with human body parts). We propose three approaches in this section to address the above mentioned three IE tasks, respectively.

### 3.1 Ontology-based Medical Term Identification

Medical term identification is often a task during patient record processing. For example, clinicians are always interested in the medical history and surgical history of patients in our project. For the following description of past medical history, the system is expected to extract three terms: *postoperative CVA*, *cholecystectomy*, and *midline hernia*.

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

---

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

Medical term identification essentially belongs to the task of named entity recognition. However, medical terms are full of synonyms and morphologic variants. It is necessary to adopt ontology (or dictionary) for high accuracy extraction of medical terms from clinical records. Because medical terms are often multi-word phrases, it is not efficient to search all combinations of sequential words in the sentence through ontology. 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 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. If a term exists in the database, we then save it and continue to look for terms after the endpoint of current term. Otherwise, we look for terms matching the next pattern from the current starting point. It is worth noting that terms in UMLS are indexed in normalized form for the purpose of efficient query. Normalization usually includes two steps: (1) getting the uninflected 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 [10].

In UMLS, each term may belong to more than one concepts and at least one semantic type is assigned to each concept. According to the possible semantic type, we can determine whether the medical term extracted is of interest or not.

In this particular project, we also need to group medical terms. For example, clinicians have particular interest in certain predefined diseases such as hypertension and we then need to identify synonyms (e.g. high blood pressure is a synonym of hypertension) of these predefined diseases. This task is simply done by lookup of synonyms in ontology.

Ontology-based method for medical term extraction achieves high precision and recall. But it still fails to retrieve a portion of terms of interest simply due to the incompleteness of ontology. We relieve the problem

by guessing some terms based on the idea that elements in parallel sentence structure should play the same role. In the following example, we recognize *splenectomy* and *gallbladder cholecystectomy* as surgeries. So we also treat *gunshot* as a surgery name though it is not defined in UMLS.

*"Gunshot wound in 1989, splenectomy in 1992, and gallbladder cholecystectomy in 1990"*

In short, ontology-based approach for medical term identification is much better than general named entity recognition approaches in terms of precision and recall. However, it is searching intensive though we adopt ordered part of speech patterns to minimize the number of term candidates.

### 3.2 Graph-based Concept Association

Concept association refers to a task that tries to find two concepts in text (usually in a sentence or a couple of consecutive sentences) which are semantically or syntactically related to each other. Most of information extraction (IE) tasks in this project are simply concept association or can be transformed to concept association.

One type of information for extraction is number such as blood pressure, pulse, age and weight of a patient. Because this project targets patients with breast cancer, clinicians also care about menarche age, number of pregnancies, and number of live births. The extraction of these numbers is equivalent to associate medical concepts (e.g. blood pressure) with numbers. Another type of information is about the association of diseases or symptoms with person (e.g. father, mother, aunt etc.) or the part of human body (e.g. right breast and left breast). The trace of family history of cancer is about the association of disease with person. The examination of breast is about the association of symptoms with part of human body.

Some IE tasks can be transformed to concept association problems. For example, clinicians are interested in the status of menopause of the patient. It is a typical classification problem. But browsing patient records, we found that if the date of last menstrual period is known, then the status of the menopause can be determined. Thus the problem is transformed to the association of medical term (last menstrual period) with date.

The whole procedure of concept association is comprised of two steps. The first is the identification of all concepts including diseases, symptoms, human body parts, persons, numbers, dates, etc. The second is the association of concepts. The extraction of persons, dates and numbers is not difficult. In fact, most NLP

development tools, such as GATE, provide the module of named entity recognition modules, which annotate above-mentioned concepts in a text with extremely high precision and recall.

The identification of medical terms (human body parts, diseases and symptoms) has two modes. One is the search of certain type of medical terms. The ontology-based method introduced in Section 3.1 can be applied to this mode. The other is the search of specific concepts. One straightforward approach for this mode is exact string matching. In order to improve the recall, we also search the synonyms and inflected variants of the target concept. 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 use WordNet and some heuristics to automatically generate them from original concepts.

Pronominal reference is also required in most cases of concept identification because the concept appears in the form of pronoun such as *it* sometimes. We use some shallow methods [3] to find the real entity the pronoun refers to.

The next step is to find concept pairs which are semantically or syntactically related to each other. The task of association is quite difficult because in the majority of cases a sentence contains more than two concepts. In the first sentence of the example below, there are four medical metrics and four numbers. In the second sentence, there are two human body parts and two symptoms.

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

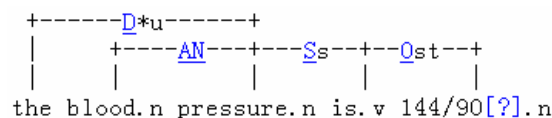
*"...There is no other mass palpable in the right breast while the left breast is free of any lesions"*

One shallow approach to the concept association is the use of linguistic patterns or heuristics. Some examples of linguistic patterns for association of medical metrics with numbers are listed below:

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

The major advantage of the pattern-based approach lies in its simplicity. However, it is difficult to exhaust all patterns since the expression of natural language is so flexible. Here we propose and implement a novel graph-based approach for concept association, which uses the linkage information produced by Link Grammar Parser [14].

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 the example shown in Figure 1, there are four links. The link between "is" and "144/90" represents a verb-object relation (denoted by notation 'O'). There is a bunch of research work that applies link grammar parser to information extraction. Madhyastha, Balakrishnan and Ramakrishnan reported the use of link grammar parser for event extraction [9] and Ding, Berleant, Xu and Fulmer applied link grammar to the extraction of biomedical interactions [4]. Both of their work reached the goal of information extraction by analyzing the meaning of important links in the sentence. Differing from their approaches, we first transform the parsed sentence to a formalism of graph and then perform concept association based on the generated graph.



**Figure 1.** An Example of a Linkage Diagram<sup>2</sup>

Suppose a node represents a word and an edge represents a link. Then the linkage diagram of a valid sentence can be viewed as a connected graph. Furthermore, each edge can be weighted against the type of link according to the application. Thus, the distance between any word pair can be calculated from the graph. Intuitively, the distance between any word pair is a good measure of their syntactic relationship. Then the task of concept association is equivalent to search the shortest node with certain semantic type for a given node in a (weighted) graph.

For some information extraction tasks, we need pay attention to the occurrence of negative words or phrases. In the following example, *left breast* is associated with the symptom of *lesions*, but because of the occurrence of negative phrase, *be free of*, they actually have no association at all.

*"...There is no other mass palpable in the right breast while the left breast is free of any lesions"*

This approach provides a generic framework for concept association. But it has several technical limitations in practice. First, link grammar parser can not parse text fragments without verb (e.g. blood pressure: 144/90). For this reason, we also implement the pattern-based approach. If the parser fails to parse

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

the sentence, the pattern approach will take the place. Second, link grammar parser is originally developed for conversational English and makes many errors while parsing text in biomedical domain, most likely due to its lack of the syntactic information of biomedical vocabulary. Third, link grammar parser can process single-word concepts but can not deal with multi-word concepts because.

Regarding the last two problems, Szolovits presents a heuristic method to augment the lexicon of link grammar parser with UMLS's specialist lexicon. We plan to adopt this technique in future version. In the current project, we instead use a simple method to relieve the problem. After medical term identification, we replace these terms in sentence with place holders and then submit the modified sentence to parser. The last example sentence will be converted to the sentence below after if our method is applied. Though link grammar parser can't recognize the meaning of these place holders, but it is able to figure out the part of speech of these holders and successfully parses the sentence.

*"There is no other symptom1 in part1 while part2 is free of any symptom2"*

In this sub-section, we introduce a graph-based approach for concept association. In comparison with pattern-based approach, it is more flexible and robust. This approach is comprised of five steps. They are in order: concept identification, pronominal reference, medical term replacement, link grammar parsing, and graph building.

### 3.3 Decision Tree based Text Classification

Text classification is another type of information extraction tasks in our project. For instance, patients fall into three classes with regard to smoking behavior: non-smoker, former smoker, or current smoker. The following texts are examples describing different smoking behaviors.

*"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 [11] 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 (uninflected 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 inflected form is indicative of classifications.

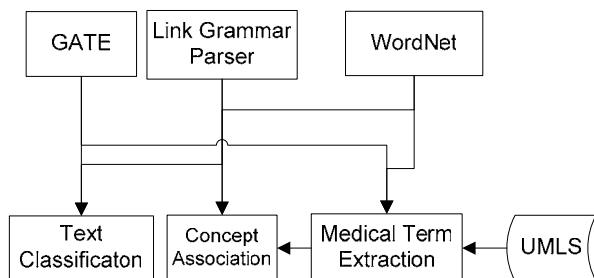
In classification of smoking behavior, we search for parts of speech — verbs, nouns, adjectives, or adverbs — that appear in any constituent part of the sentence; meanwhile, we disable the "head noun or head adjective only" option, and enable the "use of lemma" option. In classification of appearance, however, only adjectives are selected because clinicians in our case usually use adjectives to describe patients.

The above method of feature extraction works well for most text classifications in our project. But for classifications containing numeric information, the 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 use other techniques such as concept association to find new features. Still in this example, we get the average

amount of alcohol use per day of the patient using concept association and then add a Boolean feature, which is about the comparison of the average amount with some standard, to the mode. The performance of the new model is dramatically improved.

## 4. Implementation

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



**Figure 2.** System Architecture

Link Grammar Parser<sup>3</sup> is used to produce both linkage information for concept association and constituent trees for feature extraction during text classification. WordNet<sup>4</sup> is mainly used to get the lemma (uninflected form) of each surface word in a sentence. GATE<sup>5</sup> (General Architecture for Text Engineering) is used for tokenization, sentence splitting, and part of speech tagging. UMLS serves as the domain ontology for medical term identification. For the sake of efficiency, we downloaded UMLS data and installed it in a local DB2 database. The data is accessed through JDBC. We implemented the ID3-based decision tree algorithm for text classification.

## 5. Evaluation

The data set consists of fifty patient records, each of which is for a subject with breast complaints. The data

<sup>3</sup> 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>.

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

<sup>5</sup> <http://gate.ac.uk/>

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.

We use precision and recall to evaluate the 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.

Field (attribute) name	Precision (Recall)
Blood pressure	100.0%
Weight	100.0%
Pulse	100.0%
Age of menarche	100.0%
Number of pregnancies	100.0%
Age of first child	100.0%
Number of live births	100.0%
Menopause	94.0%
Palpable nodule	86.0%
Breast Mass	86.0%
Axillary Nodes	100.0%
Nipple D/C	100.0%
Family History of cancer	92.0%
The reason to visit doctor	92.0%

**Table 1.** The performance of information extraction using concept association

The extraction of fourteen attributes listed in table 1 based on the method of concept association achieves extremely high precision (recall). By examining all fifty records manually, we find that the extremely high precision is in part 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.

Classification Tasks	Precision (Recall)
Smoke behavior	92.2%
Alcohol use	89.4%
Appearance	93.7%

**Table 2.** The performance of text classification

The ID3-based decision tree is evaluated on the three classification tasks: smoking behavior, alcohol use, and appearance. 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 then calculated (see table 2).

Clinicians in our project are also interested in the medical history and surgical history of the patient. Because these attributes may contain multiple values (medical term), the precision and recall for i-th patient 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 performance of medical term extraction is listed in table 1.

Attribute Name	Precision	Recall
Predefined Past Medical History	96.7%	96.7%
Other Past Medical History	88.1%	89.4%
Predefined Past Surgical History	92.3%	94.2%
Other Past Surgical History	87.5%	92.3%

**Table 3.** The performance of medical term extraction

## 6. Conclusions

In this paper, we implemented an information extraction system that extracts a variety of information of patients with breast complaints from semi-structured patient records and good performance is achieved.

The information extraction tasks in this project can be roughly classified into three classes. The first one is the extraction of medical terms. The second is text classification. The last, also the major one, is about concept association. We propose three approaches to address those three different IE tasks, respectively.

A novel graph-based approach, which uses the parsing result of link-grammar parser, was invented for concept association, and extremely high accuracy was achieved. A simple but efficient ontology-based approach 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 perform text classification. This preliminary approach to categorical fields has, so far, proven to be quite effective.

However, the size of data set used is quite small. When more diversified writing styles are introduced

into patient records, the performance of IE may be degraded. We plan to use a larger data set to evaluate and tune our future work. Besides, link grammar parser makes many errors while parsing text in biomedical domain. We are going to relieve this problem by augmenting its lexicon with UMLS's specialist lexicon in future version. Moreover, we will try to make the system more flexible and robust.

Approaches proposed in this paper 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] Dimitrov, M., Bontcheva, K., Cunningham, H., and Maynard, D., "A Light-weight Approach to Coreference Resolution for Named Entities in Text", *Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lisbon, 2002.
- [4] Ding, J., Berleant, D., Xu, J., and Fulmer, A.W., "Extracting Biochemical Interactions from MEDLINE Using a Link Grammar Parser", *In the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03)*, 2003.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.

- [9] Madhyastha, H.V., Balakrishnan, N., and Ramakrishnan, K.R., "Event Information Extraction Using Link Grammar", *13th International Workshop on Research Issues in Data Engineering: Multi-lingual Information Management (RIDE'03)*, 2003.
- [10] Miller, G. et al, "WordNet: an On-line Lexical Database", *International Journal of Lexicography*, 1990, pp. 235-245.
- [11] Quinlan, J.R., "Induction of Decision Trees", *Machine Learning*, 1986, No.1, pp.81-106.
- [12] 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
- [13] 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.
- [14] Sleator, D. and Temperley D., "Parsing English with a Link Grammar", *Third International Workshop on Parsing Technologies*, 1993.
- [15] 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.
- [16] 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.
- [17] Soderland, S., "Learning Information Extraction rules for Semi-structured and free text", *Machine Learning*, Vol. 34, 1998, pp. 233-272.
- [18] Szolovits, P., "Adding a Medical Lexicon to an English Parser", *Proc. AMIA 2003 Annual Symposium*, 2003.

## 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, Zolof, 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.