To appear in ***The 21st Annual ACM Symposium on Applied Computing 2006, Technical tracks on Computer Applications in Health Care* (*CAHC 2006*), Dijon, France, April 23 -27, 2006.**

Approaches to Text Mining for Clinical Medical Records

Xiaohua Zhou and Hyoil Han
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
xiaohua.zhou@drexel.edu
hhan@cis.drexel.edu

Isaac Chankai, Ann Prestrud and Ari Brooks
College of Medicine
Drexel University
Philadelphia, PA 19102
{ic36, ann.prestrud}@drexelmed.edu
ari.brooks@drexelmed.edu

Abstract
Clinical medical records contain a wealth of information, largely in free-text form. Means to extract structured information from free-text records is an important research endeavor. In this paper, we describe a MEDical Information Extraction (MedIE) system that extracts and mines a variety of patient information with breast complaints from free-text clinical records. MedIE is a part of medical text mining project being conducted in Drexel University. Three approaches are proposed to solve different IE tasks and very good performance (precision and recall) was achieved. A graph-based approach which uses the parsing result of link-grammar parser was invented for relation extraction; 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 was used to perform text classification.

Categories and Subject Descriptors
I.2.7 [**Natural Language Processing**]: Text Analysis

General Terms
Experimentation

Keywords
Clinical Records, Information Extraction, Relation Extraction, Ontology

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 and efficient method to extract structured information for future data mining from free-text using information extraction techniques may greatly benefit research endeavors.

We report here on the development of a MEDical Information Extraction (MedIE) system that extracts and mines a variety of patient information with breast complaints from free-text clinical records. MedIE is a part of a large research project on breast cancer being conducted at Drexel University College of Medicine. Before researchers can conduct any analysis or mining, they must first code textual patient records and save this structured information into the database. A total of 125 separate initial consultation notes were mined by our system. Results were then compared to a medical student's independent manual processing of the same consultation notes.

The technique used in this paper is an extension to [19]. In [19], three approaches to numeric attributes filling, medical term identification, and text classification were described, and an evaluation for 13 extracting tasks on a small collection of 50 patient records was reported. We extended the graph-based approach for numeric attribute filling in [19] to a generic relation extraction technique capable of performing the majority of information extraction tasks in the project; we also solved several technical problems while using Link Grammar parser [14] to build graphs, which made the system more robust. We improved term extraction approach in [19] by extensive use of ontology and adoption of a NLP-based term prediction technique. In short, the extended MedIE system is more generic, robust, and effective in terms of knowledge extraction; the evaluation for 23 extracting tasks on a larger collection of 125 patient records is more representative and convincing.

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 extraction of the three types of information; and section 4 evaluates system performance. 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 concepts, such as pattern matching, 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, the ontology is necessary to achieve high extraction accuracy. 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 such a large database of

SAC'06, April, 23-27, 2006, Dijon, France.

medical terminology for information extraction from clinical records. In this particular project, we adopt Unified Medial Language System (UMLS)1 as the domain ontology to identify medical terms.

The 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 to prepare training examples. 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 research that applies link grammar parser to information extraction. Madhyastha et al. reported the use of link grammar parser for event extraction [9] and Ding et al. applied link grammar to extraction of biomedical interactions [4]. Both works achieved 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 a graph and then perform concept association based on the graph generated.

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]. Kuhn and 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 extracting tasks in our project can be roughly classified into three groups. The first is extraction of medical terms (e.g., past medical history and past surgical history). The second is text classification. For example, a patient can be classified as a former smoker, a current smoker, or a non-smoker. The last group is about relation between two terms (e.g. symptoms and human body parts). We propose three approaches to address these extracting tasks, respectively.

3.1 Ontology-based Term Extraction

Medical term extraction is often a task during patient record processing. For example, clinicians are always interested in the medical history and surgical history of patients. Medical term extraction 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 for high accuracy extraction of medical terms from clinical records. Medical terms are often multi-word phrases; therefore, it is not efficient to search all combinations of sequential words in the sentence through the ontology. Instead, we follow the method in [19], using part of speech patterns to generate term candidates and then checking if the candidate terms exist in the ontology.

In UMLS, each term may belong to more than one concept 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. We then need to identify synonyms (e.g. *high blood pressure* is a synonym of *hypertension*) of these predefined diseases. This task is simply completed by lookup of synonyms in ontology.

The ontology-based approach for medical term extraction achieves high precision and acceptable recall. But it still fails to retrieve a portion of terms of interest simply due to the ontology incompleteness or typo made by doctors. We relieve the problem by predicting some terms based on the idea that elements in coordinating structures should have similar semantic types. In the following example, we recognize *splenectomy* and gallbladder *cholecystectom*y as surgeries; *gunshot* then has a good chance to be a surgery name though it is not explicitly defined in UMLS.

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

The ontology-based approach for medical term extraction can achieve higher performance than those general named entity recognition approaches. However, it requires intensive searching though we adopt part of speech patterns to minimize the number of term candidates.

3.2 Graph-based Relation Extraction

Relation extraction refers to a task that finds pairs of two terms in text (usually in a sentence or a couple of consecutive sentences) that are semantically or syntactically related to each other. Most information extraction (IE) tasks in this project are relation extraction or could be transformed to relation extraction problem. One type of information for extraction is numeric attribute, such as blood pressure, pulse, age and weight of a patient. Because this project targets patients with breast cancer, clinicians are also concerned about menarche age, number of pregnancies, and number of live births. Extraction of these numbers is equivalent to associated medical concepts (e.g. blood pressure) with numeric values. Another type of information is the association of diseases or symptoms with persons (e.g. father, mother, aunt etc.) or parts of the body (e.g. right breast or left breast). For example, the trace of family history of cancer is about the association of disease with a person; examination of breast is about the association of symptoms with part of human body. Some extracting tasks could be transformed to relation extraction problems. For example, clinicians are interested in the menopausal status of the patient. This is a typical classification problem. But browsing patient records, we found that if the date of last menstrual period is known, then menopausal status can be determined. Thus the problem is transformed to the association of medical term (last menstrual period) with date.

The procedure of relation extraction is comprised of two major steps. The first step is the extraction of various terms including diseases, symptoms, human body parts, persons, numbers, dates, etc, as described in Section 3.1. Co-reference resolution is required for relation extraction because doctors may use pronoun or abbreviation to reference previous terms while writing patient records. We use a shallow method [3] to find the real entity pronouns or abbreviations refer to.

[1] http://www.nlm.nih.gov/research/umls/

The second step is to find pairs of terms that are semantically or syntactically related to each other. The judgment of semantic relation is simple because the semantic type of each term is already given during extraction and the possible relations of any two semantic types are pre-defined in the ontology. However, the determination of syntactic relation is difficult because in the majority of cases a sentence contains more than two terms. 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"

We propose a graph-based approach for the extraction of syntactic relation based on 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 nine links. The link between "is" and "144/90" represents a verb-object relation (denoted by notation 'O'). Some researchers have explored the use of link grammar in information extraction. Madhyastha et al. reported the use of link grammar parser for event extraction [9] and Ding et al. applied link grammar to the extraction of biomedical interactions [4]. Both works 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.

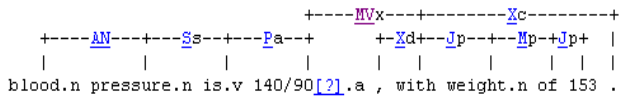

Figure 1. An Example of a Linkage Diagram[2]

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 (e.g. we penalize the links connecting two clauses). Thus, the distance between any term pair can be calculated from the graph. Intuitively, the distance between any term pair is a good measure of their syntactic relationship. Then the task of syntactic relation extraction is equivalent to search the shortest node (or the node with a distance less than the threshold [4]) with certain semantic type for a given node in a (weighted) graph.

For some extracting tasks, we need to pay attention to the occurrence of negating words or phrases. In the following example, left breast is linked 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 relation extraction, but has several technical limitations in practice. First, link grammar parser cannot parse text fragments without verbs (e.g. blood pressure: 144/90). For this reason, we also implemented a pattern-based approach. If the parser fails to parse the sentence, the pattern approach will take the place. Second, link grammar parser was originally developed for conversational English and makes many errors while parsing text in the biomedical domain, most likely due to its lack of syntactic information of biomedical vocabulary. Third, link grammar parser can process single-word terms but cannot deal with multi-word terms. Regarding the last two problems, Szolovits [18] presented a heuristic method to augment the lexicon of link grammar parser with UMLS's specialist lexicon. We plan to adopt this technique in future versions. In the current project, we used a simple method to relieve this problem. After medical term identification, we replaced these terms in sentences with placeholders and then submitted the modified sentence to parser. The last example is sentence converted to the sentence below after our method is applied. Link grammar parser cannot recognize the meaning of placeholders, but it is able to figure out the part of speech the holders represent and successfully parses the sentence.

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

In this sub-section, we introduced a graph-based approach for relation extraction. In comparison with pattern-based approach, it is more flexible and robust. This approach is comprised of following five components, term extraction, co-reference resolution, medical term replacement, link grammar parsing, and graph building.

3.3 Decision Trees 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 of the literature. Usually pattern-based semantic analysis would be performed to classify 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 employed 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 fewer features than other decision tree algorithms. (For the details, see [19])

4. Experiment

The MedIE system is implemented by Java. Link Grammar Parser is used to produce both linkage information for relation extraction and constituent trees for feature extraction during text

[2] The diagram is yielded by the online Link Grammar parser at http://www.link.cs.cmu.edu/link/.

classification. WordNet is mainly used to get the lemma (uninfected form) of each word in a sentence. GATE (General Architecture for Text Engineering) is used for tokenization and part of speech tagging. UMLS serves as the domain ontology for medical term identification. For the sake of efficiency, we downloaded the UMLS data and installed it in a local DB2 database. Data is accessed through JDBC. We implemented the ID3-based decision tree algorithm for text classification.

We evaluated our MedIE system on a collection of 125 patient records, each of which is a subject with breast complaints. The format of the patient records is same as [19]. 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.

Table 1. Result of extraction using concept association

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

We use precision and recall to evaluate the performance. Extraction of fourteen attributes listed in Table 1 based on the method of relation extraction achieves extremely high precision (recall). In [19], only the first seven tasks listed in Table 1 were performed. By examining all 125 records manually, we found that the extremely high precision is, in part, attributed to the consistent dictation style (all records were provided by the same clinician, author Ari D. Brooks, MD). If the size of the data set increases or the writing style varies, performance may be degraded.

Table 2. Result of text classification

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

The ID3-based decision tree is evaluated on three text classification tasks: smoking behavior, alcohol use, and appearance. Five-fold cross validation was 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 as testing data. We ran a five-fold cross validation ten times, each time the dataset is randomly shuffled. Average precision (recall) is then calculated (see table 2). It is worth noting that [19] only evaluated the classification task of smoking behavior.

Table 3. Result of medical term extraction

Attribute Name	Precision [ours]	Recall [ours]	Precision [19]	Recall [19]
Predefined Past Medical History	96.7%	96.7%	96.7%	96.7%
Other Past Medical History	88.1%	89.4%	76.1%	86.4%
Predefined Past Surgical history	92.3%	94.2%	77.8%	35%
Other Past Surgical History	87.5%	92.3%	62.0%	75%

Clinicians in our project are also interested in the medical and surgical history of patients. Because these attributes may contain multiple values (medical terms), the precision and recall for i-th patient are defined respectively as, $R_i = ETrue_i / TInstl_i$ and $P_i = ETrue_i / ETotal_i$. Precision and recall for the whole collection, respectively, are defined as, $R = \sum_i ETrue_i / \sum_i TInst_i$ and $P = \sum_i ETrue_i / \sum_i ETotal_i$ 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.

We revised the approach for medical term identification and achieved significant progress of precision and recall. For the details of performance improvement, please refer to Tables 3, which list the performance of medical term extraction in [19] and our experiment, respectively.

The performance improvement is mainly attributed to the extensive use of domain ontology. After medical term identification, we need to further classify terms into predefined terms or other terms. In [19], the authors failed to recognize synonyms of predefined terms. We corrected this problem, which increases the recall of predefined terms and precision of other terms. The authors of [19] simply treat all terms existing in domain ontology as medical terms of interest. Actually a small portion of extracted terms such as “history” and “human” was not expected. We filtered out these terms in the new version by additionally examining the semantic type of the term, which increases term precision. Due to the incompleteness of domain ontology (UMLS) or typo, ontology-based approach failed to extract some terms not defined in ontology. In the current version, we use the coordinating structure to predict some medical terms, which increases the recall of medical term extractions.

5. Conclusion

In this paper, we implemented a MEDical Information Extraction (MedIE) system that extracts and mines a variety of information from clinical medical records. Good performance was achieved.

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

A graph-based approach which uses the parsing result of link-grammar parser was invented for relation extraction and 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 was 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 small. When more diversified writing styles are introduced into patient records, the performance of IE may degrade. 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 the lexicon with UMLS's specialist lexicon in future versions. Moreover, we will try to make the system more flexible and robust.

Approaches proposed in this paper may offer a new means by which clinician-researchers may extract large volumes of data from patient medical records. To date, this resource is untapped, as there is no effective means to extract data. We hope to continue this work, refining our approach, to expand its utility.

6. 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.

[19] Zhou, X., Han, H., Chankai, I., Prestrud, A.A., and Brooks, A.D., "Converting Semi-structured Clinical Medical Records into Information and Knowledge", *In the International Workshop on Biomedical Data Engineering in conjunction with the 21st International Conference on Data Engineering (ICDE)*, Tokyo, Japan, April 3-4, 2005.