Preparing Clinical Text for Use in Biomedical Research

John P. Pestian, University of Cincinnati, USA

Lukasz Itert, preferred affiliation

Charlotte Andersen, preferred affiliation

Wlodzislaw Duch, preferred affiliation

ABSTRACT

Approximately 57 different types of clinical annotations construct a patient's medical record. The annotations include radiology reports, discharge summaries, and surgical and nursing notes. Hospitals typically produce millions of text-based medical records over the course of a year. These records are essential for the delivery of care, but many are underutilized or not utilized at all for clinical research. The textual data found in these annotations is a rich source of insights into aspects of clinical care and the clinical delivery system. Recent regulatory actions, however, require that, in many cases, data not obtained through informed consent or data not related to the delivery of care must be made anonymous (as referred to by regulators as harmless*), before they can be used. This article describes a practical approach with which Cincinnati Children's Hospital Medical Center (CCHMC), a large pediatric academic medical center with more than 761,000 annual patient encounters, developed open source software for making pediatric clinical text harmless without losing its rich meaning. Development of the software dealt with many of the issues that often arise in natural language processing, such as data collection, disambiguation, and data scrubbing.*

Keywords: please provide

INTRODUCTION

Hospitals typically produce millions of text-based medical records over the course of a year. These records are essential for the delivery of care but underutilized or not utilized at all for clinical research. Digitized clinical data are a rich lode of possibilities for advances in biomedical research, because, in aggregate, they contain information about the variation in the delivery and quality of care.

Inherent in such research, however, is the use of data without the patient's consent. Recognizing this problem, the United States Department of Health and Human Services (HHS) has issued rules defining Protected Health Information (PHI) as part of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (Annas, 2002). In order for researchers to access such data, either they must have the patient's consent, or, as in most retrospective cases, the data must be made harmless, and the governing board must provide a waiver.

The HHS provides guidance for making healthcare data harmless (HIPAA Standards for Privacy of Individually Identifiable Health Information: An Introduction to the Consent Debate, 2002). Data can be made harmless through three steps: (1) de-identification (i.e., the removal or modification of data fields that could identify a patient, such as name and social security number); (2) rendering the data ambiguous by ensuring that every data record in a public data set has a non-unique set of characterizing data (Berman, 2002a; Bouzelat, Quantin, & Dusserre, 1996; Quantin et al., 1998); and (3) data scrubbing (i.e., the removal or transformation of those tokens in text that can be used to identify persons or that contain information that is incriminating or otherwise private) (Berman, 2003; Sweeney, 1996). Although each of these methods has the potential to render the medical record harmless for its use by natural language processing investigators, attempts to design a fully anonymous system continue.

This article describes how Cincinnati Children's Hospital Medical Center (CCHMC), a large pediatric academic medical center with more than 761,000 pediatric patient encounters per year, has taken a practical approach to this challenge by developing, evaluating, and implementing the Encryption Broker (EB) software. The EB has a number of uses. First, it is essential for the ongoing development of a large pediatric corpus for pediatric natural language processing research and decision support (Pestian, Itert, & Duch, 2004). This corpus serves as an artificial intelligence training set for classifying text into the appropriate clinical domain, such as rheumatology or neonatology. Without the EB, these data could not be retrieved from the electronic portion of the medical records. Second, the EB ensures that research-needing text conforms to federal regulations. It does so through data disambiguation algorithms, de-identification, and data scrubbing.

The EB has another role. A key strategy of the organization is personalized medicine research that requires genomic and clinical delivery data to predict or prevent disease or to personalize treatment. This research requires substantial amounts of knowledge to be gleaned automatically from these data in real time. To do so, machine-learning systems that conceptually map the data into some ontology are required. The EB provides natural language scientists with large repositories of harmless clinical text for developing these systems.

The EB is recognized by CCHMC's Risk Management group as a tool to gather clinical text without violating HIPAA regulations. This approval is institution-specific; each institution using the EB is responsible for seeking its own internal certification. The EB essentially acts as a broker for investigators who wish to do retrospective analysis of clinical text and potentially makes it easier to receive approval for these purposes. CCHMC makes the EB software, the associated decision rules, and the related data files fully available through its Web server (http://info.cchmc.org) for academic purposes. The remaining sections of this article discuss methods and challenges for making these data harmless, CCHMC's approach, and the evaluation of this methodology.

LITERATURE REVIEW

It is beyond the scope of this article to describe fully the rich history of research in the areas of natural language processing; this review highlights those areas that have contributed to developing the conceptual approach underpinning the research presented: word sense disambiguation and data scrubbing.

WORD SENSE DISAMBIGUATION

Examining tokens in their context and determining exactly what sense is being used is the task of Word Sense Disambiguation (WSD). WSD is a difficult task and, as such, receives considerable theoretical and practical attention. To disambiguate (i.e., OR vs. operating room) requires an understanding of the surrounding tokens. In other words, "You shall know a word by the company it keeps" (Firth, 1957). There are two ways to do this. One is a supervised approach that integrates rule-based information into the semantic analysis. The other is an unsupervised stand-alone approach, where sense disambiguation is performed independent of and prior to compositional semantic analysis.

For this research, integrated rule-to-rule approach was used, because raw clinical notations are heavily packed with jargon, and unsupervised methods are traditionally used with well-formed text. Ng and Zelle (1997) note:

> *For each token to be disambiguated, the appropriate inference knowledge must be handcrafted. It is difficult to come up with a comprehensive set of the necessary disambiguation knowledge. Also, as the amount of disambiguation knowledge grows, manual maintenance and further expansion become increasingly complex. Thus, it is difficult to scale up manual knowledge acquisition to achieve wide coverage for real-world sentences.*

This summary points out the limitations and provides future research guidance. That is, since rule-to-rule WSD requires substantial effort at some point, it will be necessary to integrate this work into a stand-alone unsupervised machine learning system.

Determining the optimal window size for token analysis is another important task. The linguistic tools used for WSD can be divided into two general classes: collocation and co-occurrence. Collocation, a quantifiable position-specific relationship between two lexical items, encodes local lexical and grammatical information that often can accurately isolate a given sense (Jurafsky & Martin, 2000). In collocation, the assumption is that some tokens often are found together (e.g., emergency room or breast milk).

Co-occurrence data focus on the frequency of the same token within a particular range of tokens while ignoring its position. For example, "John's parents were in the *emergency room* while the *emergency room* physician treated John." Co-occurrence focuses on the fact that *emergency room* occurred twice. Collocation focuses on the fact that *emergency* is located next to *room*.

These tools enable selection of specific domain tokens from a larger generalized corpus (Jurafsky & Martin, 2000). This study formally uses local collocations to disambiguate terms. In particular, +/- three tokens around the target token (**t**) were analyzed. This window of tokens is referred to as a trigram. This strategy was based on previous research that notes:

> *[L]ocal collocation provides the most important source of disambiguation knowledge, although the accuracy of disambiguation achieved by the combined knowledge sources exceeds that obtained by using any one of the knowledge sources alone. That local collocation is the most predictive seems to agree with past observation that humans need a narrow window of only a few tokens to perform WSD. (Ng & Zelle, 1997)*

DATA SCRUBBING

The literature describes many forms of data scrubbing. Scientists use data-scrubbing methods to de-identify pathology data (Berman, 2003), threshold cryptographic protocols (Berman, 2002b), automate record hash coding and linkages for epidemiological follow-up data confidentiality (Quantin et al., 1998), object-oriented software components (Herting & Barnes, 1998), cryptographic framework for document objects resulting from multiparty collaborative transactions (Goh, 2000), use personal identifiers while retaining confidentiality in child abuse cases (Kruse, Ewigman, & Tremblay, 2001), and describe data hiding techniques (Chao, Hsu, & Miaou, 2002).

Although research in the area of de-identification has been active over the last few years, scholars are still undecided as to whether it is possible to fully de-identify data. For example, Sweeney and Dreiseitl conclude that most data can be re-identified by linking or matching the data to other databases or by looking at unique characteristics found in the fields and records of the database itself (Dreiseitl, Vinterbo, & Ohno-Machado, 2001; Sweeney, 1997a). Sweeny, after reviewing a number of data-scrubbing systems, concludes that removing all explicit identifiers from medical data does not guarantee anonymity; rather, complementary policies will be necessary (Sweeney, 1997b). Others, however, regard those processes as too onerous to yield any practical consideration (Fisher, Baron, DJ, Barett, & Bubolz, 1990). Until the optimal set of strategies is found, each institution must address problems with de-identification as it finds best.

METHODS

A patient's medical record is comprised of approximately 57 different types of documents (Zweigenbaum, Jacquemart, Grabar, & Habert, 2001). These documents contain both structured data (e.g., computerized order entry data) and unstructured data (e.g., clinical dictations). Some data are confidential; others are a matter of public record. Computerized or handwritten notes include birth and death records, discharge summaries, imaging reports, short problem descriptions, and letters (Friedman, 1997; Grefenstette, 1994; Sager, Friedman & Lyman, 1987; Zweigenbaum & Menelas, 1994). The content of these documents has a great deal of variation not only between the documents but also within the documents themselves (Biber & Finegan, 1994). This study concentrates on unstructured clinical text found in discharge summaries, radiology reports, surgical reports, and pathology reports.

The minimum regulatory standards for making PHI harmless require removal of up to 16 specific pieces of information (Madsen, Masys, & Miller, 2003). In the case of unstructured text, simply removing or encrypting these identifiers will disrupt the ability to understand the PHI and its meaning, thus rendering it useless for natural language processing research.

The remaining sections of this article outline the methods for collecting data, development of rules, three stages of software development, and the evaluation of the software.

DATA COLLECTION

From 2000 to 2002, CCHMC's division of Biomedical Informatics developed the Discovery System (DS), a centralized research repository (Pestian, Aronow, & Davis, 2002). The DS is populated regularly with new and updated clinical, research, and administrative data generated by the medical center. Substantial amounts of these clinical data are text from such specialties as pathology and radiology and from discharge summaries and surgical notes. The DS combined with other data are used for studying genotypic prediction of pharmacological responses and microarray expression of newborn hearing testing, sepsis onset in intensive care patients, the onset and severity of juvenile rheumatoid arthritis, quality assurance, financial reporting, and other activities.

Access to the data for research is governed by HIPAA regulations and controlled by the organization's Institutional Review Board (IRB). Prospective studies receive approval before the study begins. Access to retrospective data also must receive approval from the IRB. Requests for text that are not part of a formal research study sanctioned by the IRB are approved only after the data have been made harmless by using the EB or some other method.

DATA CLEANSING METHOD

The data-cleansing algorithm relies on two steps in order to render the unstructured clinical text harmless and preprocess it for use. The first step is to disambiguate the unstructured clinical text that is dense with jargon and acronyms. The second step is data scrubbing or de-identification. Each of these steps is described in subsequent paragraphs.

Figure 1. Trigram analysis neighborhood

Analysis Neighborhood

Notation	B_{T3}	B_{T2}	B_{T1}	τ	A_{T1}	A_{T2}	A_{T3}
Example	John	was	born	FT	with	no	complications

Where B_{tn} = tokens before the Evaluation token, E_T = evaluation token and A_{TN} equals tokens after the evaluation token.

WORD SENSE DISAMBIGUATION

"All grammars leak" (Sapir, 1921). This is because people are always stretching and bending the rules to meet their communicative needs (Manning & Schutze, 1999). It should be no surprise that extensive jargon and acronyms have leaked into clinical text. The language of clinicians, though fundamental to patient care, lacks the structure and clarity necessary for natural language analysis. For example, in a clinical text, the token FT can be an abbreviation for *full-term, fort* (as in Fort Sumter)*, feet or foot, field test*, *full-time*, or *family therapy*. Until these text data are disambiguated, there is no certainty that data scrubbing is accurate.

To resolve the ambiguities found in the text, a series of clinical disambiguation rules were made. The data were stored in the *rules.dat* file. The first step for developing these rules was to create a reference dataset that contained known ambiguous terms, clinical acronyms, and abbreviations. After developing a dataset of known acronyms and abbreviations, clinical experts reviewed the text for ambiguous terms. Ambiguous terms were added to the dataset. This review was done three times until the experts believed that most ambiguous terms were included in the dataset.

This reference dataset was then used to create a dataset of trigrams. Software was developed to extract from the all the data the trigrams for each ambiguous term. Clinical experts then reviewed these trigrams to create the disambiguation rules. Figure 1 presents a schematic of this approach. In the figure, one term, *FT*, is being evaluated by looking at the three tokens before *FT* and the three tokens after *FT*. The experts then reviewed all the trigrams and developed the disambiguation rules, using a majority/minority approach. That is, all instances of a specific term (i.e., *FT*) remain as a specific term (i.e., *FT*), unless an evaluation parameter is met. For example, one rule is *If FT* if followed by *with*; then FT = *Full-Term*.

DATA SCRUBBING

Once the data were disambiguated, they were reviewed for the presence of any of the 16 possible Protected Health Information (PHI) data elements. Limited PHI was found in the unstructured text fields. What were found were the patient and physician names and, rarely, a date of service; all other PHI was located in other structured database fields and could be eliminated by excluding those fields from the original query. Next, systematic bias was introduction into the data as a method of encryption; all female names were changed to *Jane*, all male names were changed to *John*, and all surnames were changed to *Johnson*. Table 1 provides an example of how the input data were changed.

TOKEN EVALUATION

The token evaluation criteria are based on the n-gram approach where n = the number of tokens to be evaluated before and after the token under consideration. The default value is NGRAM = 3, or a trigram. Thus,

$$\boldsymbol{\tau};|||\ ||||;\boldsymbol{\delta}$$

is the syntax to evaluate a particular token, where **τ** represents the token under consideration, and **δ** represents its replacement. In

Table 1. Output example

Before	After
Fred Thompson* is an 8 y/o AAM with a hx of asthma. He presented in the ED with a laceration on his R radius approx 3 in. long.	John Johnson is an eight-year-old African American male with a history of asthma. He presented in the emergency department with a laceration on his right radius approximately 3 inches long.

** Fred Thompson is not the patient's name.*

this case, because NGRAM = 3, the series of pipes (|) symbolize the trigram of three tokens preceding and the three tokens following **τ**; seven pipes yield six points to evaluate **τ**. Bigrams would have five pipes, and so forth. In this case, the positions are |1|2|3|4|5|6|.

Numbers 1, 2, and 3 are positions of tokens preceding **τ**; numbers 4, 5, and 6 are positions of tokens that occur after the **τ**. Consider the following sentence.

The patient stayed in OR for one hour.

Each token is assigned a position:

Patient/1/ stayed/2/ in/3/ for/4/ one/5/ hour/6/

τ = *OR* it is excluded from the position assignment.

Next is the syntax for the rule to evaluate the abbreviation *OR* based on its collocation to *IN* using one of the more than 40 predefined conditions and placed at the NGRAM position. Detailed software documentation is included with the download.

```
or;|||CONDITION(in)|||;operating room
```

If the condition is fulfilled, then the abbreviation *OR* will be replaced with *operating room* in text. If the condition is not fulfilled, then the next condition is considered. This occurs until the last condition is evaluated via an exit criterion.

A typical rule for the mentioned example could look like this:

```
or;|||IS(in)|||;operating room
or;||||IS(for)||;operating room
or;||||FINAL()|||;or
```

The first condition evaluates if the token *in* is before *operating room*. If this condition is not satisfied, the second condition is analyzed. If *OR* is followed by *for*, then *OR* is replaced with *operating room*, but if this is not true, the last condition says that the token should remain as *OR*. There are more than 40 predefined conditions (e.g., IS, PRE_NUM, POST_NUM) that can be used for testing. By default, all abbreviations are converted to lower case. Table 2 shows the pseudo code and the corresponding syntax.

EVALUATION

Evaluation of the EB consisted of randomly selecting encrypted sentences and pairing them with the original sentences. Clinical experts then reviewed these data and classified each token into one of four categories: a correct replacement, an incorrect replacement, a correct miss, and an incorrect miss. Proportions were then computed.

RESULTS

Processing scripts were written in Perl 5.0. Processing took place on a Sun Microsystems E6500, using 12 900-Mhz processors with 24 GB RAM.

DATA COLLECTION

All 2002 clinical texts were extracted from the DS. Table 3 provides the descriptive statistics for these data.

Table 2. Disambiguation rule: Pseudo-code and rule coding

<table>
<tr><th>Pseudo-Code</th><th>Rules File Coding*</th></tr>
<tr><td><b>Evaluation Token = ALL</b>
If <i>ALL</i> is all upper case and preceded by HISTORY OF, RULE OUT, RULING OUT, H/O, B-CELL, T-CELL, FOR, HIGH RISK, #REFRACTORY, PROBABLE, WITH, T-CELL, PRE-B, RELAPSED, then change <i>ALL</i> to Acute Lymphocytic Leukemia.
If <i>ALL</i> is upper case and followed by LOW RISK, then <i>ALL</i>= Acute Lymphocytic Leukemia.
#All others stay as <i>ALL</i>.</td><td>%ALL;|||PRE_ISM(out,h/o,b-cell,t-cell,for,risk,refractory,probable, with,t-cell,pre-b,relapsed)||||;Acute Lymphocytic Leukemia
%ALL;||PRE_INC_PHR(history,of)|||||;Acute Lymphocytic Leukemia
%ALL;||PRE_INC_PHR(low,risk)|||||;Acute Lymphocytic Leukemia
ALL;|||FINAL()||||;ALL</td></tr>
<tr><td><b>Evaluation Token = mm</b>
If <i>mm</i> is preceded by moist, dry, pale, sticky, tacky, then change <i>mm</i> to mucus membranes.
If <i>mm</i> is followed by moist, dry, tacky, pale, sticky, then change <i>mm</i> to mucus membranes.
If <i>mm</i> is immediately preceded by a number (i.e., 100, 7.1, 13-14, etc.), then change <i>mm</i> to millimeters.
If <i>mm</i> is followed by clinic, repair, sac, workup, surgery then change <i>mm</i> to Myelomeningocele.
If <i>mm</i> is immediately preceded by, diagnosis of, known, h/o, history of, s/p, secondary to, with, then change <i>mm</i> to Myelomeningocele
All others stay as <i>mm</i></td><td>mm;|||PRE_ISM(moist,dry,pale,sticky,tacky)||||;mucus membranes
mm;||||POST_ISM(moist,dry,tacky,pale,sticky)|||;mucus membranes
mm;|||NUM()||||;millimeters
mm;||||POST_ISM(clinic,repair,sac,workup,surgery)|||;Myelomeningocele
#ADDED:
mm;|ANY_ISM(known,vp,shunt,thoracic,secondary, spina,bifida)||||||;Myelomeningocele
#
mm;||IS(diagnosis)|IS(of)||||;Myelomeningocele
mm;|||ISM(known,h/o,s/p,with)||||;Myelomeningocele
mm;||IS(history)|IS(of)||||;Myelomeningocele
mm;||IS(secondary)|IS(to)||||;Myelomeningocele
mm;|||FINAL()||||;mm</td></tr>
</table>

**Note:_Full technical documentation is provided online at http://info.cchmc.org.*

Table 3. Descriptive statistics

Description	**Total**
Total tokens in data set	19,924,949
Total sentences in data set	1,263,271
Average tokens/sentence (standard deviation)	15.33 (9.93)
Total paragraphs in data set	173,933
Average number of sentences per paragraph (standard deviation)	7.42 (20.44)
Total unique tokens in data set	129,282
Total trigrams in data set	20,291,335
Total unique trigram in data set	5,118,035

Table 4. Evaluation descriptive statistics

Description	Total
Tokens	133,210
Sentences	10,240
Average tokens per sentence (SD)	13 (9)
Correct number of changes (% of Total Tokens)	1420 (99.2)
Incorrect number of changes (% of Total Tokens)	110 (0.08)
Correct number of non-changes (% of Total Tokens)	132,670 (98.93)
Incorrect number of non-changes (% of Total Tokens)	770 (0.58)

DISAMBIGUATION

Tokens of 915 candidate-ambiguous terms included all approved hospital acronyms, unapproved acronyms, and terms that were found. Clinical experts reviewed 715,518 trigrams that included these ambiguous terms. From this review 1,146 distinct rules for resolving ambiguity of the tokens were developed. Each rule was based on reviewing an average of 781 trigrams for a particular ambiguous token. These rules were then added to the EB's rules file to enable ambiguity resolution during data parsing.

DATA SCRUBBING

A review of the text fields found that the PHI present in the text clinical annotations was the patient's name, physician's name, and various dates. This finding made the algorithm for data scrubbing rather straightforward; by the introduction of systematic bias, data could be changed without compromising their meanings. The software's algorithm changed all male names to John, all female names to Jane, all surnames to Johnson, and all dates to 01/01/2005. This version of the software does not deal with neutral names (e.g., Pat). Future versions will.

EVALUATION RESULTS

The EB was evaluated by randomly selecting 348 records (a 0.05, 95% CI) from the original data and pairing these data with the corresponding data output from the EB. Table 4 shows the results of this comparison. A total of 10,240 (paired) sentences were reviewed by clinical experts. Ninety-eight percent of the time, the EB correctly changed a token; equally important, 99% of the time, when a token should not be changed, it was not. Of those tokens that were incorrectly changed (0.58%), a clear pattern emerged. The majority of these errors were related to ambiguous names. For example, the token *may* can mean the given name *May*, the month of *May*, or the command that *he may play sports in two weeks*. Errors in the output were found when any of the supporting files were not kept current.

SUMMARY

Protecting health information always has been a responsibility of healthcare organizations. Now that HIPAA regulations require additional levels of accountability, healthcare organizations must be creative when rendering such data harmless for research purposes. This approach shows that this is possible, but it has taken considerable effort, expense, and resources to develop and to evaluate the appropriate software. For example, to develop the first set of rules, the process includes collecting data, manually reviewing more than 700,000 trigrams to develop more than 1,000 disambiguation rules.

An important next step will be to determine the possibility of migrating from a handcrafted rules approach to rules that are made based on supervised or unsupervised machine learning algorithms. A recent paper by Liu et al. (2004) best describes this discussion: "Supervised WSD is suitable only when we have enough sense-tagged instances with at least a few dozens of instances for each sense." Here, sense-tagged refers to ambiguous tokens that have been clarified via various methods like collocation or co-occurrence. "The combination of collocations and neighboring tokens are appropriate selections for the context. For terms with biomedical unrelated senses, a large window size such as the whole paragraph should be used, while for general English words a moderate window size between four and ten should be used" (Liu, Teller, & Friedman, 2004). Thus suggesting that the optimal method by be a combination of hand-crafter rules, and machine learning.

Other questions remain unresolved. First, how generalizable are disambiguation rules? That is, is the jargon used by physicians in one part of the country or in one hospital, for that matter, different from the jargon used in another part of the country or another hospital? Second, how generalizable are disambiguation rules from the pediatric population to adult populations? While it is conjectured that there is little differences, certain differences will be inherent in the populations (i.e., adults will not be diagnosed with atrial septal defects; likewise children will not have coronary artery bypass grafts procedures). Third, how will a patient's longitudinal records be linked with this approach?

REFERENCES

Annas, G. J. (2002). Medical privacy and medical research — Judging the new federal regulations. *N Engl J Med, 346*(3), 216-220.

Berman, J. J. (2002a). Confidentiality issues for medical data miners. *Artif Intell Med, 26*(1-2), 25-36.

Berman, J. J. (2002b). Threshold protocol for the exchange of confidential medical data. *BMC Med Res Methodol*, *2*(1), 12.

Berman, J. J. (2003). Concept-match medical data scrubbing. *Arch Pathol Lab Med, 127*(6), 680-686.

Biber, D., & Finegan, E. (1994). Intra-textual variation within medical research articles. In N. Ooostdijk, & P. de Haan (Eds.), *Corpus-based research into language* (Vol. 12) (pp. 201-222). Amsterdam: Rodopi.

Bouzelat, H., Quantin, C., & Dusserre, L. (1996). Extraction and anonymity protocol of medical files. In *Proceedings of the AMIA Annual Fall Symposium* (pp. 323-327).

Chao, H. M., Hsu, C. M., & Miaou, S. G. (2002). A data-hiding technique with authentication, integration, and confidentiality for electronic patient records. *IEEE Trans Inf Technol Biomed*, *6*(1), 46-53.

Dreiseitl, S., Vinterbo, S., & Ohno-Machado, L. (2001). Disambiguation data: Extracting information from anonymized sources. In *Proceedings of the AMIA Symposium* (pp. 144-148).

Firth, J. E. (1957). A synopsis of linguistic theory 1930-1955. In F. R. Palmer (Ed.), *Selected papers of J. R. Firth 1952-1959* (pp. 1-32). London: Oxford Philogical Society.

Fisher, E. S., Baron, J., DJ, M., Barett, J., & Bubolz, T. (1990). Overcoming potential pitfalls in the use of Medicare data for epidemiologic research. *Am J Public Health, 80*(12), 1487-1490.

Friedman, C. (1997). Towards a comprehensive medical natural language processing system: Methods and issues. *Journal of the American Medical Informatics Association, 4*(Suppl), 595-599.

Goh, A. (2000). Cryptographic framework for document-objects resulting from multiparty collaborative transactions. *Stud Health Technol Inform, 77*, 1069-1073.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. London: Kluwer.

Herting, R. L., Jr., & Barnes, M. R. (1998). Large scale database scrubbing using object oriented software components. In *Proceedings of the AMIA Symposium* (pp. 508-512).

HIPAA standards for privacy of individually identifiable health information: An introduction to the consent debate. (2002). *J Health Law, 35*(3), 387-394.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.

Kruse, R. L., Ewigman, B. G., & Tremblay, G. C. (2001). The zipper: A method for using personal identifiers to link data while preserving confidentiality. *Child Abuse Negl, 25*(9), 1241-1248.

Liu, H., Teller, V., & Friedman, C. (2004). A multiaspect Comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association*.

Madsen, E., Masys, D. R., & Miller, R. A. (2003). HIPPA possums. *Journal of the American Medical Informatics Association, 10*(3), 294.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Ng, H. T., & Zelle, J. (1997). Corpus-based approach to semantic interpretation in NLP. *AI, 18*(4), 45-54.

Pestian, J., Aronow, B., & Davis, K. (2002). *Design and data collection in the discovery system*. Paper presented at the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Science.

Pestian, J. P., Itert, L., & Duch, W. (2004, May). *Development of a pediatric text-corpus for part-of-speech tagging*. Paper presented at the Intelligent Information Systems, Poland.

Quantin, C., Bouzelat, H., Allaert, F. A., Benhamiche, A. M., Faivre, J., & Dusserre, L. (1998). Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods Inf Med, 37*(3), 271-277.

Sager, N., Friedman, C., & Lyman, M. (Eds.). (1987). *Medical information processing — Computer management of narrative data*. Reading, MA: Addison Wesley.

Sapir, E. (1921). *Language: An introduction to the study of speech*. New York: Harcort Brace.

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the scrub system. In J. J. Cimino (Ed.), *Proceedings, Journal of the American Medical Informatics Association* (Vol. 1996, pp. 333-337). Washington, DC: Hanley & Belfus, Inc.

Sweeney, L. (1997a). Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium* (pp. 51-55).

Sweeney, L. (1997b). Weaving technology and policy together to maintain confidentiality. *J Law Med Ethics, 25*(2-3), 98-110, 182.

Zweigenbaum, P., Jacquemart, P., Grabar, N., & Habert, B. (2001). *Building a text corpus for representing the variety of medical language*. Paper presented at the Medinfo.

Zweigenbaum, P., & Menelas, C. (1994). Menelas: An access system for medical records using natural language. *Compt Methods Programs Biomed, 45*, 117-120.

John Pestian heads the Computational Medicine Center (www.computationalmedicine.org), an Ohio Third Frontier initiative. He is an associate professor of pediatrics at Cincinnati Children's Hospital Medical Center, University of Cincinnati. His area of research focuses on clinical natural language processing.

Lukasz Itert is a doctoral student in the Department of Informatics, Nicolaus Copernicus University in Torun, Poland. Currently, he is on a scholarship as a research assistant in the Division of Biomedical Informatics at Cincinnati Children's Hospital Medical Center, University of Cincinnati. His research interests focus on natural language processing, machine learning, and information retrieval methods in the medical domain.

Charlotte Andersen is a project manager in the Division of Biomedical Informatics at Cincinnati Children's Hospital Medical Center. She received her master's in Pediatric Nursing from the Medical College of Virginia in Richmond and has over 17 years experience in research and clinical care.

Wlodzislaw Duch heads the Department of Informatics, Nicolaus Copernicus University, Torun, Poland, and is also a visiting professor at Nanyang Technological University in Singapore (2003-2007). Currently he is the president of European Neural Network Society. To find more information about him write "Duch" in Google.