Mining Free-Text Medical Records

Daniel T. Heinze, PhD; Mark L. Morsch, MS; John Holbrook, MD
A-Life Medical, Incorporated
San Diego, California

Text mining projects can be characterized along four parameters: 1) the demands of the market in terms of target domain and specificity and depth of queries; 2) the volume and quality of text in the target domain; 3) the text mining process requirements; and 4) the quality assurance process that validates the extracted data. In this paper, we provide lessons learned and results from a large-scale commercial project using Natural Language Processing (NLP) for mining the transcriptions of dictated clinical records in a variety of medical specialties. We conclude that the current state-of-the-art in NLP is suitable for mining information of moderate content depth across a diverse collection of medical settings and specialties.

Introduction

In the United States alone, medicine is a trillion dollar per year business and generates in excess of seven hundred million clinical documents (about three terabytes) in transcribed free-text form. Viewing medicine as a business, the clinical information in the free-text records has a necessary application in producing a bill for services and facility utilization. Additionally, this information could be used to track physician performance and resource utilization. From the clinical perspective, the information in the clinical notes could be used to improve communications between multiple providers, to monitor the efficacy of alternate courses of treatment and to provide feedback and alerts relative to the course of care for a particular patient.

Although the Electronic Medical Record (EMR) has been a major goal in Health Information Management (HIM) for decades, the success of such systems has been seriously limited due to the relative inaccessibility of the information in free-text clinical documentation. Attempts to change the documentation habits of physicians have not had significant success largely due to the increased time and inconvenience associated with using computer interfaces that require formatted input. Further, numerous consultations with practicing physicians have taught us that there is a basic inability of fully structured systems to represent many of the nuances that make each case unique.

A-Life Medical, Inc., has developed LifeCode®, an NLP engine to abstract/code medical documents.[1,2,3,4] A detailed, system-level description of LifeCode® can be found in Heinze et al.[1,2]

LifeCode® provides both linguistic competence and medical knowledge and analysis to:

1. Use NLP to extract from a free-text clinical note...
   a) the patient demographics (name, age, gender, etc),
   b) the patient's chief complaint,
   c) the history of the present illness (duration, severity, time of onset, circumstances of medical relevance, related signs and symptoms, location of the injury/illness, context of onset, etc.),
   d) the medical history of the patient and (as applicable) patient's family,
   e) relevant social history (use of tobacco, alcohol and drugs, living arrangements, etc.),
   f) the nature and extent of the physical examination performed by the physician,
   g) the nature and extent of old records consulted, professional consultations and medical tests performed by the physician,
   h) the final diagnoses, potentially also including possible and ruled-out diagnoses,
   i) the course of treatment including surgical procedures, drug therapy and monitoring levels,
   j) the severity of the patient's condition in terms of the physician's stated conclusions and as measured by co-morbidities and the nature and course of treatment, and
   k) the disposition of the patient at the end of the clinical encounter with the physician.
2. Use domain knowledge to determine from the extracted information...
   a) the most specific version of each diagnosis and procedure,
   b) the risk to the patient presented by the medical condition and treatment,
   c) the complexity of the medical decision making for the physician,
   d) the level of service provided by the physician, and
   e) the information that can be directly reported and that which may require validation, augmentation or correction.

1067—5027/01/$5.00

Demands of the Target Market

Several applications are enabled with NLP technology – coding and billing, population of a structured electronic medical record from clinical free-text documentation and feeding a resource utilization monitoring system. All are essentially text mining operations with similar demands in terms of source documentation and specificity and depth of information required. Another application of pure text mining relates to epidemiology and outcomes analysis as viewed by the pharmaceutical industry.

The pharmaceutical industry and clinicians desires to track the epidemiology of diseases and conditions in which the use of drug therapy is crucial to patient treatment. Additionally, it is desirable to track the course of treatment and outcome for individual patients. Moreover, there are additional dimensions to the text mining problem. Text mining queries are not limited to areas that have been the subject of ongoing medical studies – i.e. the results must come from the raw clinical documentation. Tracking needs to be performed in real-time on a daily basis. In addition, new criteria may be added to the basic search at any time for both retrospective and forward-looking events. In essence, this requires a relatively deep analysis of original source documentation. The reasons for wanting such information can range from a desire to determine which courses or treatment are effective for particular conditions per patient group to wanting to know where the latest outbreak of community acquired pneumonia is taking shape so that a sales force can be first to market.

Table 1 lists a general set of issues related to a text-mining project for acute myocardial infarction. The actual set of queries decomposes these general queries into specific signs and symptoms, diseases, medical procedures and medications.

The list is similar to that for other medical conditions in that it can be divided into several categories of information. Demographic information is important first to identify and track individual patients across multiple medical encounters. For multiple visits to the same facility, medical record and account numbers and the date of service can generally track a patient. It may be necessary, however, to track patients across visits to multiple facilities. In this case, it is important that the source of clinical documentation has good coverage for the facilities in a selected geographic region. Also, due to the ambiguities of names, it becomes necessary to use personal information such as gender, age and even medical history to track individual patients. History of the medical condition is required to understand how the condition originated or was first noticed, what the patient did in response to it, the response to prior treatment (whether performed by the patient or by a medical professional), the characterization of the condition in terms of signs and symptoms, the progress of the condition, etc. The subjective responses of the patient to medical inquiry regarding their current and recent condition must be evaluated. The physical examination by the medical staff, the resultant diagnoses, the methods of treatment and the final disposition of the patient for each medical encounter must all be quantified.

a.	Identify survival rate during hospitalization.
b.	Group by presence of risk factors, e.g. family history, hyperlipidemia, diabetes millitus, cigarette smoking, prior myocardial infarction.
c.	Identify presence of co-morbidities, e.g. valvular disease, chronic obstructive pulmonary disease.
d.	Group by presenting symptoms, e.g. chest pain, dyspnea, arm/leg pain, jaw pain.
e.	Group by location of infarction: anterior, anteriolateral, inferior, right ventricular.
f.	Group by type of infarction: transmural vs. non-transmural.
g.	Group by duration of hospitalization.
h.	Identify 5 most common types of arrhythmias present during hospitalization.
i.	Group by use of thrombolytics.
j.	Group by use of aspirin after hospital presentation; where possible, identify time interval between hospital presentation and administration of aspirin.
k.	Identify patients undergoing acute cardiac catheterization versus later.
l.	Group catheterized patients by major anatomic abnormalities: left main, LAD, left circumflex disease: and significant one vessel, two vessel or three or more vessel disease.
m.	Group by discharge medications: aspirin, beta-blockers, anticoagulation.

Table 1: Acute MI Text Mining Requirements

An information coding system is required to provide the structure for data mining. If the queries were confined to examples such as "identify all patients who had an acute MI", a simple binary encoding scheme would work. The follow-up query "and identify each patients co-morbidities" means that a far richer, and preferably portable, coding system is required. For many of the target issues of medical text mining, there are extensive coding systems. ICD-9-CM and CPT are the most common and accessible coding systems for diseases and procedures. Augmenting these codes with modifying factors such as severity, methodology, etc. can enrich each of these code sets. Some uniformity for modifier coding can be achieved by using the coding schemes from the Unified Medical Language System (UMLS), SNOMED-RT, Read Codes, etc. Although these are less universally implemented, they provide a coding mechanism that is not purely ad hoc. The Center for Medicare and Medicaid Services (CMS – formerly HCFA) has defined various "counting" schemes for elements of the history, exam and medical

decision making. The National Drug Code (NDC) provides a unique identifier code for every legally manufactured and marketed pharmacological substance. Given the looming HIPAA portability regulations[5], it can be hoped that coding systems will make significant advances in terms of coverage, consistency and acceptability over the next several years. However, until that time, there will continue to be a significant level of ad hoc nature in medical coding.

Text Quality and System Requirements

Clinical documentation is huge in quantity, frequently below par in grammaticality and is characterized by multiple levels of subdomain vocabulary.

As we have noted, the domain of clinical documentation is extensive, consisting of more than three terabytes a year of transcriptions of dictated clinical encounter notes. A-Life Medical's data mining partner, MedQuist, alone produces about 20% of this text with near 100% coverage in numerous key geographic regions. Transcribed medical records are produced for virtually all encounters in referred medicine, e.g. radiology, surgical pathology, etc. Transcription is also nearly universal for acute care medicine (at least at the discharge summary level) and surgery. Transcribed notes comprise the majority of the clinical notes for specialty medicine. Recent advances in automated speech recognition[6] and pressures from both the government and major medical carriers promise to make electronic clinical notes the norm for general and family medicine.

The common medical parlance of a dictated clinical note is considerably different from the language of medical texts, references and scholarly publications. Some physicians are very terse. A few are excessively verbose. In almost all cases, incomplete sentences abound. Frequently, the location of a statement within the physical format of the document is information bearing.

Medical vocabulary and usage is a multi-tiered system. At the top level, there is reference terminology. It has been developed over time in an attempt to make medical communications succinct and unambiguous. At the second level, there is the medical vulgate – the language commonly used in clinical documentation in various levels of mixture with the medical reference terminology. As the frequency of a medical condition or procedure increases, the probability of vulgarization and divergent local usage also increases. Conversely, if a condition or procedure is extremely rare, the tendency is for the language to become idiosyncratic and governed by the whim of those few practitioners who have a command of the science. The reference terminology is relatively straightforward to incorporate into a computational system. The idiosyncratic terminology is infrequent enough that it can be "almost known"; a concept that is exploited by LifeCode®.[2] The vulgate requires more attention. We have approached it by attempting to fit the system to the basic characteristics of the physician speakers.

As a case study, this paper cannot go into the complete analysis of the domain characteristics and NLP approach. Some examples will have to suffice. We have noted that our approach to language is cognitivist.[7] That is to say, we believe that the basic facilities and expressions of human language are rooted in the more basic cognitive processes associated with the perceptual abilities. This is a controversial position in theoretical linguistics, but continuing developments in cognitive science and the success that we have had with our technology have given us confidence that this position is correct. Of course, this paper is not the appropriate venue in which to argue the point. It is of note, however, because clinical medicine is very oriented toward sensory perception in its linguistic expression. We would attribute this to its long history and tradition that far predates laboratory medicine. Clinical medicine is firmly rooted in what the physician and/or patient can feel, see, hear, smell and taste. These basic sensory tests are crossed with the perception of the passage of time. Even at the intersection of clinical and laboratory medicine, the perceptual metaphor takes over. A white blood count noted by the clinical physician only in numerical terms is more often than not normal, whereas an abnormal count is interpreted in terms of spatial perception metaphors as being either low or elevated. As a response to this observation (and based on our basic cognitivist presuppositions), LifeCode® consists of a wide array of independent and semi-independent processors that roughly correspond to questions like "where is the problem located?", "how long ago did it happen?", "how often does it happen?", "how big is it?", "how many are there?", "how hot is it?", "what color is it?", "how much does it hurt?", "what type of pain is it?", etc.

Additionally, the system must be able to deal with reference or attribution – i.e. "who said it?" Although each document is ostensibly the report of a particular physician describing an encounter with a particular patient, there are a number of variants on the theme. Firstly, it may not be the examining physician who is actually dictating the record. In many cases a resident, a physician assistant or even another physician may dictate the note on behalf of the examining physician. This third party may also have participated in the medical process and have described both their own

work and that of the physician. Secondly, the physician will frequently refer to portions of the work-up that were performed by other staff and specialists. This may be in terms of describing the work of others or in terms of reporting just their conclusions. Thirdly, numerous people who are in some way related to the patient and the current medical condition may be referenced. The patient's social and family medical history will be reported in terms of medical conditions and social practices of others. Also, it is common that the patient may have been brought for medical attention by another person who will provide part of the report that is related by the physician in the history.

Beyond keeping straight who said what and about whom they said it, it is important to maintain a belief framework. Firstly, in terms of whether the person making a claim is believed to be authoritative in the domain of the claim. For example, a patient can make the claim of having substernal chest pain and such a claim can be taken as authoritative. However, if the patient claims that they read a book on chest pain and are convinced that they have rheumatic fever and an enlarged heart, that claim cannot be taken as authoritative. The physician, based on a physical examination and appropriate tests, can make such a claim authoritatively. Secondly, belief is based on how sure the person reporting is of the report. Medicine is, at times, a matter of probabilities. Injuries are, more often than not, relatively straightforward to diagnosis with certainty. Many illnesses, on the other hand, are difficult to diagnosis with certainty and may be reported as possible or probable diagnoses. As to the patient's report of symptoms, it may be difficult to exactly specify just what type of pain it is or to rank just how much harder it is to breathe today as compared to a week ago. Thirdly, belief is based on just how sure the reader, whether human or machine, is that they have correctly assessed or abstracted the physician's note.

System Accuracy

Given that the system is competent in these performance areas, quality must be monitored and controlled. We have used two methods of system validation. The first is comparison of LifeCode® results against those of experienced human coders. The second is comparison of the output of LifeCode® against statistical norms.

In the arena of coding for billing purposes, both we and our customers have done multiple comparison studies showing that LifeCode® is at least as accurate as human coders on the documents that it marks as requiring no human review.[3,8] About 70% of the documents processed by LifeCode® are so indicated as requiring no human review. Most of the documents that are sent for human review are so routed because of issues related to billing regulations. In this regard, over-sampling is required due to the extremely high liability associated with billing errors. Of the documents that are sent for human review, 60+% pass review with no changes. Of the documents that require changes, the changes affect only about 5% of the information that has been extracted for that document.

For text mining, we have additionally employed statistical analysis. If, for example, we are mining information related to patients with acute myocardial infarction (MI), it is an easy statistical check to see if the percentage of patients that LifeCode® is finding with acute MI matches our prior knowledge of the number of patients who have acute MI's. The same can be done for various co-morbidities, treatments, etc. For the more detailed information that is not available for validation purposes, manual quality assurance, e.g. sequential sampling is needed until such time as statistical norms can be established. Once quality is assured and norms are established, a continued sequential sampling (now at a minimal level) assures reliability of the data and deviations from the expected values beyond the allowable limits indicate validated trends.

Results

A test case consisting of 53,656 medical notes from across the range of ambulatory and acute care clinical settings and specialties at four major university medical centers and one private medical center was run. Three disease profiles were mined. These were the acute myocardial infarction profile discussed earlier and one each related to asthma and gallbladder disease. Each profile required the mining of demographic information, primary diseases, co-morbidities, medications, medical and/or surgical interventions, and outcomes. The tests were designed so that crosschecks were performed to validate results. For example, the severity of an MI or an asthma attack was to be assessed both by the stated assessment of severity as given by the clinician and also by the type and number of treatments administered and the need for follow-up.

Setup for the test included reviewing a sample of the transcription headers for any unusual formatting styles relative to extracting the header demographics, adding study specific groupings of medications, diagnoses and procedures to the LifeCode® knowledgebases, and creating batches of files for the test runs. Setup time was approximately 24 person hours.

Processing was performed in parallel on six Pentium II and Pentium III PCs during off-peak hours. Processing time averages ~10 seconds per document on a 550 MHz processor with 256 MB RAM.

A summary of results is presented in Table 2. Manual sampling of the results validated a ~99% accuracy level. To achieve this high accuracy, certain specificity constraints were relaxed as follows. Several items requested in the profiles were determined, even before the test, to be beyond the ability of LifeCode® in its current form to extract. In some cases this was because accurate determination of the fact often required the unification of information given across multiple documents – e.g. the determination of whether a gallbladder removal was emergency or scheduled. In other cases this was because the information was not reliably reported in the clinical documentation at hand – e.g. how soon after the onset of an MI the patient was administered aspirin. Finally, some information would have required a development effort beyond the scope allowed for in the test – e.g. determining not just that a laparoscopic technique was used for a gallbladder removal, but also which of the four specific techniques defined by CPT was used. Of the 39 major categories of information specified in the three profiles, there were two urgency-related items, one technique-related item and one timing-related item that could be extracted with an accepted level of accuracy only by generalizing the profile requirements for those items.

Number of documents	53,656
Number of sources	5
Query profiles and number / percent of encounters identified for each.	Acute myocardial infarction – 854 / 1.6%
	Acute exacerbation of asthma – 1695 / 3.2%
	Gallbladder disease – 372 / 0.7%
Number of queries	39
Accuracy	~99%
Processing platform	550 MHz Pentium III
Processing time	~10 sec / document

Table 2: Summary of Results

Discussion

Medical text mining is characterized by market requirements for very precise information at a moderately deep level. The volume of available text is measured in terabytes per year and growing. This text is frequently idiosyncratic and is grammatically, more often than not, of an ill-formed nature. In response to the linguistic and semantic characteristics of the domain, the LifeCode® extraction process is at its core cognitivist. It employs a wide array of independent and semi-independent agents each of which operates to recognize patterns of usage and meaning that are loosely defined in terms of basic perceptual capabilities such as perception of size, duration, intensity, physical structure, etc. Multiple results are computed in parallel, are compared against one another in terms of likelihood and saliency, and, in the end, derive their interpretation as they map onto a framework of usage defined by the problem at hand. The quality of the overall system is assured by means of statistically controlled measures such as sequential sampling and by statistical comparison to historically established expectations.

Measurable results from a full range of clinical settings and across diverse disease and treatment profiles show that LifeCode® is reliable and accurate. Further, the low cost of abstracting with LifeCode® makes it suitable for mining clinical documents that would be prohibitively expensive using human abstracters. Suitable applications include both electronic medical record (EMR) population and text mining for trend and outcomes analysis.

References

1. Heinze, DT, ML Morsch, RE Sheffer, et.al. A natural language processing system for medical coding and data mining, *AAAI - Twelfth Innovative Applications of Artificial Intelligence Conference.* July 2000.
2. Heinze, DT, ML Morsch, RE Sheffer, et.al. LifeCode: A deployed application for automated medical coding. *AI Magazine.* vol. 22, no. 2, pp. 76-88. Summer 2001.
3. Morris, WC, DT Heinze, HR Warner Jr., et.al. Assessing the accuracy of an automated coding system in emergency medicine. *Proceedings – AMIA Annual Symposium.* November 2000.
4. Heinze, DT, J Holbrook. The need for natural language processing. *Advance for Health Information Executives.* November 2000.
5. Amatayakul, M. The race to standardize medical record information, *MD Computing.* November – December 2000.
6. Holbrook, J. Speech recognition – has it finally arrived? *Advance for Health Information Executives.* May 2000.
7. Langacker, RW. Explanation in cognitive linguistics and cognitive grammar. *Conference on the Nature of Explanation in Linguistic Theory.* University of California at San Diego – Dept of Linguistics. December 3-5, 1999.
8. Warner, Jr. HR. Autocoding medical records using natural language processing. Submitted to *Advance.*