

A-Life Medical I2B2 NLP Smoking Challenge

System Architecture & Methodology

Ronald E. Sheffer, Brian C. Potter, Ph.D., Mark L. Morsch, Daniel T. Heinze, Ph.D.
A-Life Medical, Inc., San Diego, California

ABSTRACT

We describe the architecture of LifeCode® (A-Life Medical, Inc.), a natural language processing system for free-text clinical information extraction, and our methodology in applying LifeCode® to the I2B2 NLP smoking challenge.

INTRODUCTION

LifeCode® is a natural language processing (NLP) and medical coding expert system that extracts and normalizes demographic and clinical information from free-text clinical records. First commercially available in 1998, LifeCode® currently encompasses products covering the domains of radiology, pathology and emergency medicine. Today, LifeCode® processes over 2 million medical reports per month, for over 500 sites, representing physician management groups, major medical centers, and medical billing companies.

In this paper, we describe the basic architecture of the LifeCode® system, and the methodology of applying CM-Extractor, a second-generation LifeCode® engine designed for quality measure abstraction, to the I2B2 First Shared Task for Challenges in NLP for Clinical Data smoking challenge.

ARCHITECTURE

LifeCode® is a predominantly symbolic natural language processing system that relies on morphological, syntactic, semantic and pragmatic analysis to extract and synthesize concepts from free-text medical documents. Figure 1 provides a summary overview of the LifeCode® system. A detailed system description is available elsewhere.¹

The NLP extraction engine and medical expert system modules, written largely in C++, are driven by a domain specific knowledge base. The knowledge base, abstracted away from the source code, contains the core medical vocabulary concepts, domain specific concepts such as CPT procedures and ICD9-CM diagnoses in a billing application, and concept logic. Concepts are stored with a representation similar to feature vectors which are used to map symbolic phrases to standard or proprietary codes. Concept logic provides general

and domain specific means of refining the set of concepts initially identified for a text.

When dealing with a single sentence, LifeCode® references the knowledge base an average of 50,000 times (ranging from several thousand to several million). A large table storing partial results during the vector analysis allows these calculations to be performed in typically less than one second per page of text.

The NLP module of the system is comprised of four components: document segmenter, lexical analyzer, phrase parser, and concept matcher.

The document segmenter delimits and categorizes the content of a medical note based on the meanings of any section headings within that note. Heading meanings are determined by comparison, using a flexible pattern-matching scheme, to a set of possible heading definitions specified in the knowledge base. This process places each portion of a note in a broad context, with such context preserved and accessible through subsequent stages of LifeCode® processing.

The lexical analyzer module is a series of processors designed to transform the text into a string of symbols consistent with the vocabulary of the knowledge base specifications. This functionality includes acronym expansion, morphological reduction and a variety of specialized parsers for additional input analysis and normalization.

The phrase parser employs bottom-up syntactic analysis to chunk the input into phrases. This parsing is highly tolerant of incorrect grammar and unknown words. The resulting text chunks range in size from two to three words up to a complete sentence, corresponding roughly to the granularity of concept definitions within the knowledge base.

The concept matcher uses vector analysis to assign meanings, represented as knowledge base concept labels, to each phrase. A second evaluation phase after the initial vector difference computation is used to refine the matches. This includes the use of anatomy, medication, and microbiology concept hierarchies and synonym lists to improve the chances of a match. Syntactic heuristics may be applied at this point to join and redistribute words from two or more consecutive phrases and compute the meaning for the combined phrase.

Once NLP extraction has identified an initial set of concepts for a document, the medical expert system module applies general and specialty-specific

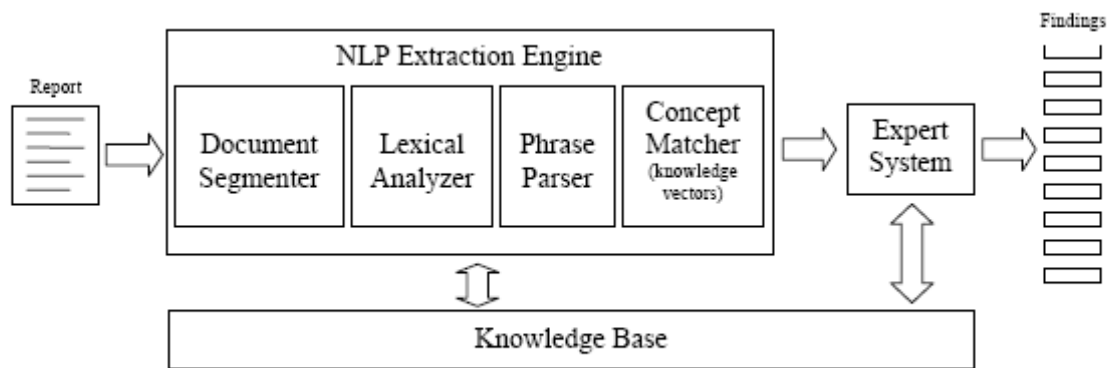


Figure 1. LifeCode® Architecture. The NLP Extraction Engine breaks reports up into concept-level phrases and matches them to known concepts using knowledge vectors.

logic to refine that set. This logic is used to resolve ambiguous concepts, eliminate redundant information, combine component concepts into a single concept, and implement application specific business rules. For example, logic may remove general concepts in the presence of related but more specific concepts, associate an exam type referenced in the ordered section of a document with an anatomical location referenced in the body of the document, or verify that essential sub-components of a procedure are present for billing purposes. The expert system logic is developed, in consultation with medical coding experts, through analysis of medical documents and published guidelines.

METHODOLOGY

For this task, we selected A-Life Medical's CM-Extractor, a second-generation NLP and clinical abstraction system designed specifically for use with JCAHO Core Measures in a multiple document inpatient record application.^{2,3} This system includes a module for detecting patient smoking status, which we adjusted for purposes of the I2B2 challenge. Adjustments were related to distinctions in smoking status, types of documents handled, sections analyzed, certainty of data, and phrases found in the I2B2 training data.

I2B2 makes more and different distinctions in smoking status than JCAHO. The JCAHO Core Measures only distinguish between "smoker", "non-smoker" and "unable to determine". A "smoker" is an individual who has smoked within the past year, while a "non-smoker" is an individual who, as indicated explicitly, has not smoked within the past year. An individual who only smokes cigars or pipes is not considered a "smoker". In I2B2, however, four categories delimit the JCAHO "smoker" and "non-smoker" categories. An I2B2 "current smoker" is an individual who has smoked within the last year, while an I2B2 "past smoker" is an individual who smoked

previously but quit more than a year prior to the discharge date. I2B2 includes the category "smoker" for individuals who smoked at some point, but where the time frame is unclear, and the category "non-smoker" is reserved for individuals who have never smoked. Finally, unlike JCAHO, I2B2 does not distinguish pipes and cigars from cigarettes. Any tobacco smoker may be assigned to one of the I2B2 smoking categories. To accommodate these categorization differences between JCAHO and I2B2 in our CM-Extractor system, we adjusted the time frames for our smoking distinctions and added the category of "smoker" for cases in which the time frame was unclear. In addition, we eliminated the JCAHO distinction between cigar and pipe versus cigarette smoking per I2B2 guidelines.

The CM-Extractor system was designed to handle a full range of document types relevant to an inpatient medical record. Such document types include history and physical reports, emergency reports, nurse's notes, discharge summaries, progress reports, consultations, discharge summaries, lab reports, orders, etc. As the I2B2 smoking challenge was limited to discharge summaries, many of the processing distinctions CM-Extractor would typically manage across document types could safely be discarded.

On the other hand, dealing with a large cross section of documents to determine a single smoking status, as required in the CM-Extractor project, allowed for a certain amount of redundancy and predictability. Smoking history, for example, is typically documented in history and physical reports, emergency medical documents, discharge summaries, and/or consultations. Because of this, evidence for smoking status could be restricted in CM-Extractor to only the most reliable locations for this type of information, namely "Past, Family and Social History" or "History of Present Illness" sections. For the I2B2 challenge, we discovered that we couldn't rely on data redundancy. Many documents in the

training set had a single mention of the patient's smoking status, and it was often in non-canonical sections. We expanded the sections examined by CM-Extractor for smoking status to include both diagnosis type sections and "course" or treatment sections.

Because CM-Extractor was designed to process inpatient records involving multiple documents and types of documents, dictated by multiple physicians and at different times, system output includes a certainty score (high, medium high, medium, and low) to indicate confidence in each extracted quality measure. Contradictions across documents, for example, lower the confidence score of a reported measure. I2B2 presented the CM-Extractor system with the challenge of reporting one and only one best hypothesis on each document. To this end, we ranked the likelihood of statuses and scored each document based on the highest ranking measure returned.

CM-Extractor, as with the other LifeCode® systems, is driven by a domain specific knowledge base. With respect to "smoking" terminology, we used the I2B2 training data to expand the types of concepts and phrases included in the CM-Extractor knowledge base and/or refine existing concepts and phrases.

TRAINING DATA ANALYSIS

In this section, we discuss the results of applying our engine to the I2B2 Challenge training data. We discuss our initial results, changes made to the system, problems encountered in adapting to the challenge, and our final results on the training data.

For our initial implementation, we changed our smoker categorizations to match the I2B2 guidelines, narrowed the document types handled, and restricted our certainty to only report the highest ranking smoking status. In the first test run, our engine's score matched the I2B2 score in 355 out of the 399 (89%) training documents. This is largely due to our previous development efforts with the JCAHO Core Measures data.

Over subsequent development cycles, we analyzed cases in which our engine was incorrect, revised the knowledge base to better match the I2B2 training data, and reran to verify improvement. We also allowed our system to mine document sections previously ignored for smoking data, such as diagnosis and course sections. And based on the training data, we incorporated some inferential references to smoking status. For example, while "quit smoking" implies that the patient does not currently smoke but did at some previous time, "attempting to quit" implies strongly that the patient

is still smoking. On document 85, for example, we encountered "Please attempt to quit smoking." in an "Additional Comments" section. The physician doesn't explicitly say that the patient does or does not smoke. Similarly, on document 406 we found "we emphasized that it would be important to stop smoking". On the basis of examples such as these, we added linguistic variants of "request to stop smoking" as strong indicators of "current smoking" status.

There were two problematic cases related to the way our engine makes use of document section information. First, because "history of" statements are not codeable from a diagnosis section in inpatient visits (or are handled differently), our engine is currently set to ignore such statements in a diagnosis section. We found two cases of "history of tobacco use" present only in a diagnosis section, but chose not to change our engine processing to handle these cases. Second, one of the training documents had no sections. Our system is set to route such documents to a human reviewer, and again we chose not to change the current processing.

We note a few inconsistencies in the training data, mostly related to time statements. From the social history on document 596, we find:

"He has a sixty to seventy five pack year smoking history and drinks alcohol approximately one time per week."

This was considered "smoker" in the training data, but in the absence of any information about the patient quitting, this seems to be a clear statement of "current smoker" status.

Several documents were classified as "past smoker" but there is no clear time frame given and "past smoker" should be accompanied by an explicit or implicit time reference indicating that the patient quit more than a year ago.

On document 621, the "history of present illness" section states:

"Briefly this a 65-year-old gentleman with a history significant for lung cancer status post complete right pneumonectomy in 1997 , COPD , past tobacco use..."

This was classified as "past smoker" in the training data, but, again, with no time frame for quitting, this seems to be an example of "smoker" status.

Similarly, on document 654, in the "social history" section, we find:

"He smoked a pack and a half for five years. He is not a current smoker."

This is classified as “past smoker” but with no time frame given (and assuming the physician’s definition of “current smoker” may be different from the I2B2 definition), this is better considered just “smoker”.

While we disagreed with the classification of these documents, we didn’t discard them from the training set or ignore them in our internal scoring evaluation. We chose not to adapt our system to these examples, however, as we thought that might negatively impact our handling of other documents.

In our final run against the training data, our engine’s score matched the I2B2 score in 371 out of 399 documents (93%). We then halted development, and subsequently downloaded and ran the 104 I2B2 test documents through the system.

Based on our initial evaluation of the test results (absent an I2B2 scoring key), we were generally pleased with our system’s performance. We did note a few obvious mistakes due to novel formatting not present in the training data. While we believe our system is linguistically robust, it does rely on the standard formatting and segmentation generally found on medical documents.

TEST RESULTS

The I2B2 confusion matrix for the LifeCode® engine’s challenge performance, and the precision, recall and f-measure scores (rounded to 4 decimal places) for all categories are reproduced below.

A-Life

U	N	P	S	C	I2B2
63	0	0	0	0	u
1	15	0	0	0	n
1	2	4	1	3	p
0	2	0	0	1	s
0	1	3	0	7	c

	Precision	Recall	F-Meas
Unknown	0.9690	1.0000	0.9840
Non-Smoker	0.7500	0.9375	0.8333
Past-Smoker	0.5714	0.3636	0.4444
Smoker	0.0000	0.0000	0.0000
Current-Smoker	0.6364	0.6364	0.6364

The weighted f-measure for our system was 0.8388. The mean across all participants in the I2B2 challenge was 0.7957.

Overall we agreed with the I2B2 scoring of the documents. Systematic issues we recognized in our review included cases in which our engine failed to recognize novel formatting (e.g., documents #41 and #176) and cases in which our engine’s

production design prevented information extraction, such as the “history of” issue with diagnosis sections (e.g., document #194).

We also found 5 cases where the I2B2 data was miscategorized, and the A-Life categorization was correct (e.g., documents #56, #660, #685, #906 and #233). While we recognize that human coders may disagree on the coding of a document,⁴ we felt that the coding for these five cases was clear. Three of these cases are shown below. Document 685, for example, contained:

“SOCIAL HISTORY :
Widowed since 1972, no tobacco, no alcohol, lives alone.”

A-Life scored this as “non-smoker” while I2B2 considered this “past-smoker”. Similarly, the “social history” section of document 906 included:

“No alcohol use and quit tobacco greater than 25 years ago with a 10-pack year smoking history.”

A-Life scored this as “past-smoker” while I2B2 scored “current-smoker”. And document 660 included:

“He is a heavy smoker and drinks 2-3 shots per day at times.”

A-Life scored this as “current-smoker” and I2B2 scored this as “past-smoker”.

CONCLUSION

For this test, our team adapted the LifeCode® engine, an NLP system in commercial use for medical coding. Our previous development experience in JCAHO core measures abstraction provided the foundation for extraction of smoking status. The total development effort for this project, including document review and performance evaluation, was relatively small, about 20 hours. The engine performed with reasonable consistency across the training and test data, with a weighted f-measure of 0.8388 for the test data. Just like in medical coding, a ‘gold standard’ for this task is an elusive goal because of ambiguity in the sources and disagreements among evaluators. However, we do believe that these types of evaluations provide meaningful benchmarks.

REFERENCES

1. Heinze, D.; Morsch, M.; Sheffer, R.; Jimmink, M.; Jennings, M.; Morris, W.; and Morsch, A. 2001. LifeCode: A Deployed Application for Automated Medical Coding. *AI Magazine* 22(2): 76-88.
2. Morsch, M.; Vengco, J.; Sheffer, R.; and Heinze, D.. CM-Extractor: An Application for Automating Medical Quality Measures Abstraction in a Hospital Setting. Proceedings of the Eighteenth Conference on Innovative Application of Artificial Intelligence; 2006 July 16-20; Boston, Massachusetts.
3. Joint Commission on Accreditation of Healthcare Organizations. Specification Manual for National Implementation of Hospital Core Measures Version 2.0. Available from: <http://www.jointcommission.org/PerformanceMeasurement/PerformanceMeasurement/Current+NHQM+Manual.htm>. Accessed September 18, 2006.
4. Morris, W.; Heinze, D.; Warner, H.; Primack, A.; Morsch, A.; Sheffer, R.; Jennings, M.; Morsch, M.; and Jimmink, M.. Assessing the Accuracy of an Automated Coding System in Emergency Medicine. Proceedings of the AMIA 2000 Annual Symposium; 2000 November; Los Angeles, California.