Leading the next generation of coding technology

Natural language processing with Optum™ LifeCode®

By Mark Morsch, Vice President of Technology

Computer-assisted coding (CAC) is a health care application whose time has come. Over the past 15 years, it has grown from a vague idea to a sophisticated solution used in hundreds of hospitals, surgery centers, and clinics across the United States.

As defined by the American Health Information Management Association (AHIMA), computer-assisted coding is "the use of computer software that automatically generates a set of medical codes for review, validation, and use based upon clinical documentation[1]."

A driving force behind CAC's progress is a technology known as natural language processing (NLP). NLP is the intelligence engine that scans and analyzes clinical documentation, then recommends codes for assigning to a clinical case. Some NLP technologies in use today show great promise. Others are likely to be abandoned in the next few years as less effective. In this paper, we will discuss the five major NLP technologies, the benefits exhibited by each, and the distinctiveness of the only patented NLP technology for health care: Optum™ LifeCode® NLP.

NLP: A technology category with a variety of approaches

NLP as a term identifies a set of technologies and approaches, each of which vary in terms of their effectiveness. Most NLP technologies available today for CAC fall into one of five methods:

- **Medical dictionary matching:** Matches individual words or groups of words found within the documentation to standard terminology from a medical dictionary. For words that match, the text is typically highlighted and validated by the coder.
- **Pattern matching:** Extends the capabilities of medical dictionary matching, by coordinating terms with specific patterns of text that describe a diagnosis or a procedure.
- **Statistical:** Gathers information from a large, pre-coded sample of documents to train algorithms based upon word and pattern distributions.
- **Symbolic rules:** Analyzes language using rules or lexicons[2], identifying the elements of language with symbols that can be manipulated by the system.

[1] AHIMA e-HIM™ Work Group on Computer-Assisted Coding. "Delving into Computer-assisted Coding" (AHIMA Practice Brief). Journal of AHIMA 75, no.10 (Nov-Dec 2004): 48A-H (with web extras).
[2] Liddy, E.D. "Natural Language Processing." Encyclopedia of Library and Information Science, 2nd Ed. New York City. Marcel Decker, Inc., 2003.

- **Symbolic rules and statistical components:** Utilizes both symbolic NLP and a mathematical model of linguistics, including semantics (levels of language that contribute to meaning) and pragmatics (applying domain knowledge to recognize information in the correct context).

To understand how these methods differ, we first need to define the standard measurements of NLP accuracy:

- **Precision** measures the number of accurate results *compared to total results*. Higher rates of precision mean lower false positives.
- **Recall** measures the number of accurate results *compared to the potential number of accurate results*. Higher rates of recall mean lower false negatives (or missed codes).

Medical dictionary matching NLP typically produces the highest number of medical terms highlighted as potential codes. Precision of medical dictionary matching is very low, due to the low number of accurate hits compared to the high number of total hits. This method does little to enhance coder productivity, since coders are left to sift through many false positives to find accurate codes.

Pattern matching NLP is more precise than medical dictionary matching, returning fewer false positives. But because it can't analyze the meaning and subtleties of language, it has somewhat lower recall than medical dictionary matching. Neither medical dictionary nor pattern matching techniques include the intelligence to apply coding guidelines to their analysis.

Statistical NLP relies on a large sample of documents where the meaning of the language has already been matched to accurate results. Only then can the training algorithm start to perform its analysis, form word-type distributions, and derive correlations between input and results that the statistical NLP can apply.

Statistical NLP systems can often be trained quickly to a moderate level of recall and precision, but high performance can be limited by the availability of a highly accurate training sample and the need to have a large number of examples of each specific coding scenario.

Symbolic rules NLP uses inference rules to interpret meaning from text, and therefore yields high precision rates (fewer false positives). Symbolic rules introduce more sophisticated techniques for analyzing medical language based upon parsing phrases and sentences. Experts in linguistics construct symbolic rules based upon parts of speech and standard English syntax. A medical condition or procedure is recognized when one or more rules successfully match a portion of the clinical documentation. Symbolic rules support more advanced language recognition, but become very difficult to maintain for large code sets like ICD-9-CM and ICD-10-CM/PCS.

Optum's LifeCode: unique approach to NLP

LifeCode NLP engine combines the strengths of symbolic rules with statistical components—a patented NLP method. LifeCode's unique technology has earned two U.S. patents.

LifeCode has sophisticated inference rules that allow it to "understand" how documentation relates to coding rules, enabling it to, among other things, correctly assign combination codes, recognize related symptoms, or differentiate personal versus family history. LifeCode integrates its symbolic analysis with a knowledgebase that consists of more than 10 million medical facts, which allows for consistent interpretation of clinical content. LifeCode presents coders with diagnosis and procedure codes that are more complete and accurate, based on their high degree of recall and precision.

LifeCode is the only patented NLP technology on the market today. The original patent—secured in 2005—describes "vector processing," LifeCode's mathematical model for isolating, comparing, and assigning different facts from clinical documentation to build a contextual framework.

Optum was awarded another LifeCode patent in 2011, which describes "mere-parsing," LifeCode's method for determining meaning from free text. Parsing is defined as the syntactic analysis of words to determine grammatical structure. Mere-parsing is the process by which LifeCode assigns meaning not merely to single phrases within a sentence, but also to a combination of related phrases from throughout the documentation.

These unique methods of organizing and extracting meaning from clinical documentation are the key ingredients that make LifeCode one of the most effective NLP technologies available today. Vector processing and mere-parsing lead to the higher degrees of precision and recall, quickly providing coders with accurate and compliant results. And better results lead to coder productivity that clinicians are expecting from CAC.

Vector processing

The 2005 LifeCode patent is a general patent that covers the full LifeCode architecture, and it describes the different language processing techniques and algorithms that make LifeCode effective.

LifeCode analyzes the medical record from multiple aspects, understanding the document bit by bit, isolating and extracting specific facts out of the document, usually in the form of noun phrases. Then it compares and assigns meaning to those facts using a technique we term "vector processing."

Vector processing is a mathematical model by which LifeCode compares its knowledgebase, using trigonometric calculations,

with the facts that were extracted from the medical record. The result is the ability of the engine to "recognize" information. It addresses the recall-related weaknesses of NLP approaches that can't analyze meaning. Instead of matching on specific words or phrases that must occur in specific order and with a very specific expression, vector processing allows LifeCode to compare facts with a broad knowledgebase, giving clients much more flexibility in the way they document cases.

Vector processing paired with symbolic rules also allows LifeCode to take into account the context of the case. Medical coders rely on many different pieces of information to form their interpretation of medical records. Such information includes the state of the patient, previous conditions, and complications and/or co-morbidities. Going beyond the patient, other contextual information could include the type of encounter, the type of document being read, the section of document, and the author of this document. All these layers provide meaning that most NLP technologies can't recognize. Vector processing with symbolic rules, however, enables LifeCode to place the specific facts that get extracted by the engine in the full and proper context.

Mere-parsing

Vector processing assigns meaning to words in a sequence, but meaning can also be found in non-sequential terms and phrases. LifeCode has the capability to parse information in ways that allow it to not just rely on single noun phrases, but to associate non-sequential terms and thus recognize more complex English expression. We call the ability to recognize such associations "mere-parsing," and it is the technology that is covered by the patent awarded to Optum in 2011.

For example, if a patient experiences pain in multiple locations, a physician may not document the pain symptoms sequentially. One paragraph may describe pain in the abdomen, while another paragraph may describe pain in the lower right flank. LifeCode can associate pain in both locations to a specific diagnosis code. This capability was covered by the 2005 LifeCode patent; the 2011 mere-parsing patent enhances this associative capability.

Mere-parsing will make LifeCode even more effective when physicians use ICD-10-CM/PCS, which includes eight times the codes as ICD-9-CM. The ICD-10 code sets place high value on granular details such as laterality and specific anatomy. In NLP technologies without the associative capabilities provided by mere-parsing, the increased documentation granularity required by ICD-10 could result in more false positive coding results. More false positives will lead to coders spending more time finding the right code, compromising productivity. With mere-parsing technology, LifeCode can effectively and quickly capture and interpret to ICD-10's high level of detail.

Another specific problem in natural language processing is negation, or the structure by which a statement is declared as false or denied. For example, the noun phrase "evidence of breast cancer" is negated when we simply add the word "no" to the beginning of the phrase: "no evidence of breast cancer." To appropriately assign a condition from clinical documentation, NLP engines must be able to account for negation. But a simple phrase such as "no evidence of breast cancer" can cause less sophisticated NLP engines to proffer false positive codes and give coders more results to sort through to find the right code(s). More complex forms of negation, such as "patient does not exhibit evidence of breast cancer," or "mammogram showed the patient free of breast cancer," make negation even trickier.

LifeCode leverages its mere-parsing methodology to handle negation by associating negative modifiers with the appropriate noun phrases. Thus, "no evidence of breast cancer" results in no "breast cancer" related codes. That's important for organizations trying to improve coder productivity rates because mere-parsing limits the number of false positives that coders must sift through.

Optum Computer-Assisted Coding and LifeCode Advantages

Technology

The patented LifeCode NLP engine provides an unmatched level of technical sophistication for processing the clinical language of medical records. In addition, the software that utilizes LifeCode provides technical sophistication with the goal of providing a highly usable front end and manageable back end. On the front end, Optum Computer-Assisted Coding (CAC) improves coder workflow and ease of use. On the back end, Optum CAC is easily managed due to a hosted architecture, single-engine configuration, robust interface capabilities, and flexible implementation timeline.

Enhanced workflow

In an automated coding environment, computer-assisted coding can simplify front end workflow by consolidating the documentation necessary for coding into one location, eliminating the need for coders to log in to as many as a half dozen systems, including the hospital information system, imaging systems, transcription, lab systems, and more.

Implementation timeline

Optum CAC solutions have been implemented at large and small provider practices with multiple combinations of information systems. We have used that client experience to refine an implementation timeline that requires just one to three weeks for the professional version of Optum CAC. The short implementation can be partially attributed to LifeCode's NLP, which requires no

"training" in physician documentation patterns. In comparison, statistical and pattern-matching NLP technologies, which do require training, can take anywhere from nine to 12 months to finalize.

Documentation improvement: As improved documentation becomes a linchpin for improving quality and revenue, NLP can be used to find holes in current documentation practices and identify specific charts with documentation deficiencies.

CAC and NLP: Technologies whose time has come

Computer-assisted coding has seen significant progress over the past 15 years, and that progress will continue—as will NLP technologies such as LifeCode. CAC applications such as Optum CAC will continue to add capabilities that will make it essential to the transformation of professional health care organizations.

The current state of health care is making health care organizations take notice of CAC and NLP. Successful organizations understand that improving their coding will improve their revenue, and that the investment made in CAC will yield far better returns. Stakeholders see that the move from ICD-9-CM to ICD-10-CM/PCS will be much easier to manage with NLP technology that can analyze hundreds of pages of clinical documentation in seconds. And in Optum CAC with LifeCode NLP, many organizations see a technology that leads the industry in terms of improving productivity, accuracy, and ROI.

About Optum

Ingenix and A-Life Medical are now OptumInsight™, part of Optum™—a leading health services business. Optum is an information and technology-enabled health services business platform serving the broad health marketplace, including care providers, plan sponsors, life sciences companies, and consumers. Its business units—OptumInsight, OptumHealth™, and OptumRx™—employ more than 30,000 people worldwide who are committed to enabling Sustainable Health Communities. OptumInsight specializes in improving the performance of the health system by providing analytics, technology, and consulting services, and is a leading provider of computer-assisted coding products and services for the health care industry.

To learn more about CAC, NLP, and Optum solutions

Call: 800.765.6705
Email: inform@optum.com
Visit: www.optuminsight.com/CACpro

OPTUM™ **www.optum.com**

13625 Technology Drive, Eden Prairie, MN 55344