THE NATIONAL LAW JOURNAL

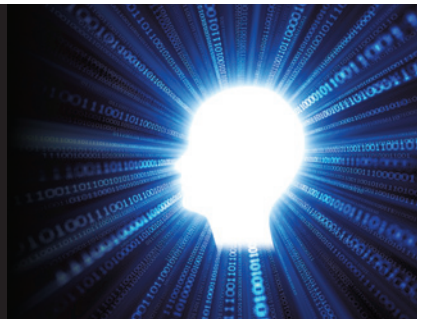
SEPTEMBER 1, 2014

Online Feature

An ALM Publication

PREDICTIVE CODING | A SPECIAL REPORT

It's been more than two years since a federal magistrate publicly endorsed the use of predictive coding in electronic discovery, and even though many lawyers still haven't heard of the technology, most of the heat has gone from the debate. Whether you call it predictive coding, technology-assisted review or computer-assisted review, the term connotes use of computer algorithms to search for dispositive evidence. In this special report, we asked litigation experts to take stock and examine some of the techniques out there.



ISTOCKPHOTO/BLACKJACK3D

Yes, Predictive Coding Works in Non-Western Languages

BY JOHN TREDENNICK

A recent U.S. Department of Justice memorandum questioned the effectiveness of using technology-assisted review with non-English documents. The fact is that, done properly, such reviews can be just as effective for non-English as it is for English documents. This is true even for the so-called "CJK languages" — Asian languages including Chinese, Japanese and Korean. Although these languages do not use standard English-language delimiters such as spaces and punctuation, they are nonetheless candidates for the successful use of technology-assisted review.

The DOJ memorandum, published on March 26, addresses the use of technology-assisted review (TAR) by the antitrust division. The author, Tracy Greer, senior litigation counsel for electronic discovery, acknowledges that TAR "offers the promise of reducing the costs" for parties responding to a DOJ second request in a proposed merger or acquisition.

Even so, Greer questions whether TAR is effective with non-English documents. "In investigations in which TAR has been employed, we have not been entirely satisfied that the TAR process works effectively with foreign- and mixed-language documents," she writes. While the division "would be open to discussion" about using TAR in such cases, she adds, it is not ready to adopt it as a standard procedure.

This is an important issue, not just for antitrust but for litigation and regulatory matters across the board. As the world gets flatter, legal matters increasingly encompass documents in multiple languages. Greer notes this in the antitrust context, writing, "As the division's investigations touch more international companies, we have seen a substantial increase in foreign-language productions."

Equally true is that discovery costs are skyrocketing as data volumes soar. Review is the costliest component of the process, accounting for more than 70 percent of discovery costs.

Cutting review costs has been the principle motivator behind the growing use of TAR, which provides a statistical basis to cut review time by half or more in many cases. Its use gained momentum in 2012, when federal and state courts first recognized the legal validity of the process.

But if TAR is of uncertain effectiveness for non-English documents, then its usefulness would be severely constrained. In a legal world in which parties and disputes routinely span borders and continents, TAR needs to work for languages other than English — and especially for Asian languages, with so much U.S. commerce flowing to and from there.

To be fair, the DOJ is not alone in questioning TAR's effectiveness for non-English documents. Many industry professionals share that doubt. They perceive TAR as a process that involves "understanding" documents. If the documents are in a language the system does not understand, then TAR cannot be effective, they reason.

Of course, computers don't actually "understand" anything (so far, at least). TAR programs simply catalog the words in documents and apply mathematical algorithms to identify relationships among them. To be more precise, we call what they recognize "tokens," because often the fragments are not even words, but numbers, acronyms, misspellings or even gibberish.

The question, then, is whether computers can recognize tokens (words or otherwise) when they appear in other languages. The simple answer is yes. If the documents are processed properly, TAR can be just

as effective for non-English as it is for English documents.

THE IMPORTANCE OF TOKENIZATION

There are a variety of TAR systems on the market, but they share one characteristic: They are computer programs. They do not understand English or the meaning of documents. They simply analyze words algorithmically according to their frequency to identify relevant documents. Lawyers train the system by marking documents as relevant or irrelevant. Based on those markings, the software analyzes the words according to their frequency, proximity and other factors to rank documents by relevance.

Nothing about TAR requires that it know English or the meaning of documents or even that it know what a word is. All it needs to do is to recognize the "tokens" — groupings of letters and characters — and analyze their relationships.

A word about tokenization: Computers do not actually search documents, they search indexes. When they process documents for search, they extract all the words and create an index. Even Google works this way, using huge indexes of words. That is how search works so quickly.

How does a computer identify what constitutes a word? It looks for series of characters separated by spaces or punctuation marks. But because not every group of characters is an actual word, information retrieval scientists call these groupings "tokens" and the act of identifying them as "tokenization."

For non-English documents — particularly for Asian documents — therein lies the rub. Certain languages, including Chinese and Japanese, do not delineate words with spaces or Western punctuation. Their characters run through the line break, often with no spaces at all.

It is true that many early English-language search systems could not tokenize Asian text. However, advanced search systems were designed with special tokenization engines capable of indexing Asian and other languages that do not follow Western conventions.

Similarly, early TAR systems focused on English-language documents and could not process Asian text. Now, however, advanced TAR systems include a text tokenizer to properly handle these languages. These systems analyze Chinese and Japanese documents just as effectively as they do English documents.

A CASE STUDY TO PROVE THE POINT

Consider an actual case we handled not long ago. A major U.S. law firm faced review of a set of mixed Japanese and English documents. Seeking to cut both the cost and time of the review, it wanted to use TAR on the Japanese documents, but was concerned about whether the process would be effective.

We worked with this firm first to tokenize the Japanese documents before beginning the TAR process. We used a method of tokenization that extracts the Japanese text from the documents and then uses language-identification software to break it into words and phrases that the TAR engine can identify.

Once we completed the tokenization, we could begin the TAR process. In this case, senior lawyers from the firm reviewed 500 documents to create a reference set to be used by the system for its analysis. Next, they reviewed a sample set of 600 documents, marking them relevant or nonrelevant. These documents then were used to train the system so it could distinguish between likely relevant and likely nonrelevant documents.

After training the TAR system, we directed it to rank the remainder of the documents for relevance. The system was able to identify 97 percent of likely relevant documents and use its ranking process to place them at the front of the review queue.

Further sampling verified that the first 48 percent of the ranking included all but 3 percent of the likely relevant documents. This low percentage in the remaining 52 percent suggested that these documents did not need to be reviewed.

Thus, by tokenizing the Japanese documents before beginning the TAR process, the law firm was able to target its review toward the documents most likely to be relevant. The firm reduced the total number of documents it needed to review or translate by more than half, thus saving roughly half the cost and time the review otherwise would have required.

The DOJ is not alone in its concern that TAR may not be effective for analyzing non-English documents. However, the fact is that computers are language-blind. They see only what they have been programmed to recognize as words, regardless of the language of those words.

With the proper technology, TAR can be used with any language — even difficult Asian languages — and produce results every bit as effective as with English. In an age of skyrocketing volumes of multilanguage data, TAR translates to savings that everyone can understand.

John Tredennick is a former trial attorney and founder and chief executive officer of Catalyst Repository Systems. He was editor-in-chief of the book "Winning with Computers: Trial Practice in the Twenty-First Century." John Tredennick can be reached at jtredennick@catalystsecure.com.