Japanese Language Expertise and Predictive Ranking Cut Review Set by 52%

Facing review of a set of mixed Japanese and English language documents, a major U.S. law firm wanted to use Predictive Ranking technology on the Japanese documents. If Predictive Ranking could identify the likely relevant Japanese documents, the cost and time required for review and translation would be significantly reduced, enabling the firm to meet client-dictated budget constraints and court-mandated discovery deadlines.

Knowing that Predictive Ranking is often ineffective when used with Japanese documents, the firm came to Catalyst for assistance. By using a method that first extracts and tokenizes the Japanese text before performing Predictive Ranking, Catalyst enabled the law firm to rank the Japanese documents for review so that the first 48% contained 98% of all that were relevant. That cut by over half the number of documents the law firm would have to review.

Predictive Ranking Alone Stumbles on Japanese Documents

The law firm received delivery of a data set for review consisting of a substantial number of documents of various file types. Of those, a third were in Japanese. The law firm wanted to apply Predictive Ranking to the Japanese documents to identify the most-likely relevant among them and prioritize them for review.

Had the documents been in English, this would have been simple enough. For Japanese documents, however, the textual analysis required for Predictive Ranking is difficult to perform. This is because Japanese does not use spaces and punctuation in the way that English does. For this reason, Japanese documents present a challenge to predictive coding engines that often leads to sub-par results.

Catalyst Tokenization Optimizes Predictive Ranking Results

Catalyst’s solution to this problem was to extract the Japanese data before beginning the Predictive Ranking process and tokenize it. Catalyst’s proprietary method of tokenization—also called segmentation—increases the effectiveness of the Predictive Ranking engine by breaking the text into words and phrases that the engine can identify.

Japanese text does not use spaces and punctuation in the way that English does, which presents a challenge to predictive coding engines.
To achieve this, we loaded the Japanese documents into our review platform. As we loaded the documents, we performed language detection and extracted the Japanese text. Then, using our proprietary technology and methods, we tokenized the text so the system would be able to analyze the Japanese words and phrases.

With tokenization complete, we could begin the Predictive Ranking process. First, 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 non-relevant. These documents were then used to train the system so it could distinguish between likely relevant and likely non-relevant documents and use that information for ranking. After the initial review, and based on the training set, we directed the system to rank the remaining documents for relevance. The results were compelling:

- Our sampling process suggested that 41% of the documents were relevant or potentially relevant. While this is higher than the norm, this custodian was known to have a high percentage of relevant documents. The system was able to identify a high percentage of likely relevant documents (98%) and place them at the front of the review queue through its ranking process. As a result, the review team would need to review only about half of the total document population (48%) to cover the bulk of the likely relevant documents.

- The remaining portion of the documents (52%) contained a small percentage of likely relevant documents. The review team reviewed a random sample from this portion and found only 3% were likely relevant. This low percentage suggested that these documents did not need to be reviewed, thus saving the cost of reviewing over half the documents.

- By applying our tokenization technology before beginning the Predictive Ranking process, we enabled our client to target its review toward the most-likely relevant documents and to reduce the total number of documents that needed to be reviewed or translated by more than half. That meant substantial savings for the client and a quicker completion of the review.

**Multi-Language Expertise Enhances Effectiveness of TAR**

Technology-assisted review often stumbles at the challenge of analyzing Japanese-language text. Catalyst’s process for first extracting and tokenizing text overcomes this challenge and makes the Predictive Ranking process far more effective for Japanese documents. For our law firm client, that meant a reduction of the Japanese documents requiring review of more than half. With fewer Japanese documents to review and translate, the firm achieved its goal of cutting the cost and time review would require.