Technology-Assisted Review (TAR)

for Asian Languages

LLM LIQUID LITIGATION MANAGEMENT, INC. | Logical Solution. Legendary Service.

ON THE SAME PAGE: TAR FOR ASIAN LANGUAGES

Advances in technology have contributed to a more fluid, seemingly border-free environment for international business. As a result, corporate lawsuits have also gone global. More than ever, counsel is likely to encounter non-English documents, specifically Asian-language, during international litigation or compliance matters.

Given the complexity of Asian languages — structural differences, context and cultural distinctions — native speakers can provide insight that is impossible to achieve through any other means. Document review by native speakers can come at a high price, however — whether financial, time or complex logistics. By applying technology-assisted review (TAR) to the review of Asian-language documents, the expertise of one reviewer can benefit many, translating into a world of time and cost savings for counsel.

By applying TAR to the review of Asian-language documents, the expertise of one reviewer can benefit many, translating into a world of savings for counsel.

TAR BENEFITS

Consider the general benefits of TAR when handling exponentially growing data sets.[1]

- First-pass tool: significantly speeds up access to important documents
- Narrows the scope of e-discovery and refines collection precision
- Serves as a compliance tool by detecting potential risk patterns

TAR is language agnostic, so these benefits also apply when implementing TAR in cases involving foreign-language documents, including Asian-language. Other benefits include avoiding the additional costs associated with international reviews (one native-language speaker versus several) and reducing risk (not relying on error-prone document translations or the complexity of large offshore-review teams).

TAR is language agnostic, so these benefits also apply when implementing TAR in cases involving foreign-language documents, including Asian-language.

A WORD ABOUT CJK LANGUAGES

While there are a multitude of Asian languages, written documents typically contain CJK languages — Chinese, Japanese and Korean. As expected, CJK languages have stark structural differences from English.

- Chinese and Japanese do not use spaces.[2]
- Japanese has four alphabets: kanji, hiragana, katakana and romaji.[3]
- Characters have multiple meanings.
- Korean uses spaces between words but follows a non-Western style.[4]

In addition to these structural differences, context can dramatically influence the content and cultural distinctions can have an impact on interpretation. By examining Chinese and Japanese more closely, we see how native reviewers can provide insight into such distinctions, and how TAR can capitalize on reviewers' knowledge.

China has a number of different regional dialects. Foreign proper nouns are usually transliterated into Chinese by using written characters for their phonetic value, rather than what they mean. Each Chinese-speaking region may — depending upon where the author grew up — transliterate the same name very differently. Compounding this is the Chinese use of acronyms, which doesn't follow defined rules.

The written Japanese language, while not as extensive as the different forms of Chinese, is one of the most complicated languages. The principal reason, mentioned earlier, is that four different alphabets can appear in a single sentence — indeed, within the very same sentence. The accuracy of a search using these characters depends entirely upon the accuracy of the technique used to extract the text from the documents.

Another factor to be aware of is that Chinese characters used in Japanese writing rarely mean the same thing as those characters in Chinese writing. And yet another is that two words, written completely differently, can mean the same thing depending upon their context (for example, rice for eating versus rice for science).

As with Chinese, transliteration of personal names can be very tricky because they may have a variety of pronunciations. Usually, personal names in Japanese are written using Chinese phonetic characters. If a name has been transliterated into English, the variations in pronunciation can be problematic.

While traditional search-term filtering or machine translation may not catch such CJK variations and complexities, native reviewers often can.

THE IMPORTANCE OF EXTRACTION AND TOKENIZATION

While traditional search-term filtering or machine translation may not catch such CJK variations and complexities, native reviewers often can.

When dealing with Asian-language documents in e-discovery, there is an emphasis on properly processing and extracting the data. Implementing TAR doesn't change this emphasis.

Prior to beginning the tagging process in TAR, data needs to be properly processed to extract the documents from multiple locations, including email. Email clients like Eudora, Lotus Notes, Thunderbird and Becky! are typically used in CJK-language countries. If the processing software cannot recognize and change character sets on the fly, so that Chinese, Japanese and Roman characters are recognized, a significant chunk of potentially relevant data is going to be missing from the extracted document text. A premier TAR software provider will not only be able to recognize the file structure, but also adjust to it.

When some file types do not have extractable text, such as images, optical character recognition (OCR) must be employed. It is imperative that the service provider counsel uses can accurately capture OCR files with Asian characters. Whether characters are extracted from their native form or OCR-ed, they must be saved in a Unicode encoding. By being saved in Unicode, they can properly express the range of characters present across all languages of the world. Unicode provides a unique number to every character — no matter the platform, program or language.[5] As a result, data can be transported through different systems without corruption or data loss.

A premier TAR software provider will not only be able to recognize the file structure, but also adjust to it.

Not only must data be extracted properly, but also the tools the platform uses, whether search indexes or analytics features like TAR, must break the content up correctly. This is known as tokenization.

When a computer indexes an English-language document, it "sees" everything between spaces or punctuation as a "token." In a traditional search, these tokens are catalogued in an index file, which contains a list of every indexed document in which that token appears. When foreign-language documents are extracted then tokenized, typically there are no spaces between words in many CJK languages. Tokenization software customized to each language relies on pattern matching so that tokens can be extracted to build the search index.

TAR and other analytics like concept search and clustering similarly require the same atomized elements of the text or tokens in order to perform their specialized form of indexing.

TAR FOR CJK: THE TECHNICAL FACTOR

With the basics of data extraction and tokenization addressed, it's on to the TAR process.

To effectively apply TAR to CJK-language documents, counsel must use a savvy software. Many search and analytics platforms rely on a static linguistic reference such as a dictionary and/or thesaurus to index documents. TAR technology, like that in LLM, Inc.'s Liquid Lit Manager™, is language agnostic. At no point does any linguistic reference point feed the technology with semantic knowledge.

TAR training begins by ingesting documents and their individual tokens, then it runs mathematical algorithms to build special matrices or indexes that correlate various terms and documents. These correlations link together conceptually similar terms and, on a larger scale, documents. The key is that the concept similarity derives from the content of the documents themselves. Practically, this highlights two advantages:

At no point does the tool have a specific language orientation. English is not its base language, and Liquid Lit Manager's TAR, for example, does not need additional linguistic and semantic plug-ins or modules to index documents in languages other than English, including Chinese, Japanese and Korean. Again, this means the technology is truly language agnostic. Thus, one advantage is that the technology can accurately find documents with similar context regardless of the language. This is of the utmost benefit when dealing with documents that have multiple languages. Conceptual correlations between terms at the heart of TAR categorization offer a second advantage by taking it a step beyond the capabilities of pure keyword search with Unicode capabilities.

Thus, one advantage is that the technology can accurately find documents with similar context regardless of the language.

TAR FOR CJK: THE HUMAN FACTOR

As discussed, the underlying TAR algorithms can accurately identify documents with similar content and context regardless of the language. TAR, however, is a process. Reviewer decisions train the TAR system that then applies those decisions to other documents based on the algorithms. This implies that high-quality human input is key to propagate decisions accurately via the TAR algorithms. To make great use of this power to amplify decisions with TAR, there is really no substitute for a native speaker of the CJK language performing the document tagging or categorization; preferably somebody also familiar with the specific industry at the core of the litigation. As the native speaker tags documents, his or her knowledge of language nuances, use of slang, cultural meaning and context provides high-quality input to the TAR workflow.

To make great use of this power to amplify decisions with TAR, there is really no substitute for a native speaker of the CJK language performing the document tagging or categorization ...

REAL-WORLD APPLICATION

One of LLM, Inc.'s clients was involved with a suit with a substantial number of Asian-language documents making up a large percentage of the total document population. The firm had a single native speaker at the office and needed to use a more costly offshore team of native speakers to complete a linear review. Given the cost and tight deadline, the firm decided to use TAR for a prioritized review.

The firm identified key custodians with a high proportion of emails containing Asian languages. Their in-office Asian language attorney performed a TAR review. Responsive documents were then expedited to the offshore team for further review, enabling them to meet their deadlines.

- Three custodian sets of emails: 100K documents, 5K reviewed, 22% identified as responsive (20K), sent for further review, recall of 94%, precision of 98.3%
- Additional custodian of 50K docs, 4K reviewed, 17% identified as responsive, recall 94%, precision 91%

CHOOSING WISELY: TAR PROVIDER AND SOFTWARE

Does the provider have a proven track record, especially with multilingual documents?

When deciding which TAR provider and software to select for TAR for foreign-language needs, consider the following qualifications.

1. Does the provider have a proven track record, especially with multilingual documents? Without relevant expertise, TAR quality can suffer and cost can increase. It's important to research potential providers and learn of their experience and case involvement.
2. Does the provider employee a proven technology or algorithm?
3. How user-friendly is the software and does it offer tools that provide transparency throughout the process? For example, LLM, Inc. offers a TAR dashboard that informs the user what decisions are being made.

When selecting a technology that has TAR foreign-language capabilities, be sure those language capabilities extend across the platform as the review and case progress. This is particularly important around the search functionality. Tools should have the ability to search multiple languages in a single document without the user having to define a specific language. All hits should be accurately highlighted, regardless of the language.

CONCLUSION

Without relevant expertise, TAR quality can suffer and cost can increase.

As the amount of data continues to increase exponentially and litigation and compliance matters travel to Asia, TAR is the ticket for counsel. With a proven service provider and software, the proven technology tool can effectively leverage the CJK-language expertise of one reviewer for the benefit of many. Not only can counsel reap valuable time and cost savings, but also feel confident about the review's future defensibility.

REFERENCES

[1] http://www.theediscoveryblog.com/category/technology-assisted-review-2/.
[2] http://en.wikipedia.org/wiki/Space_%28punctuation%29.
[3] http://support.apple.com/kb/TA39044?vie.
[4] http://www.koreanwikiproject.com/wiki/Word_spacing.
[5] http://www.unicode.org/standard/WhatIsUnicode.html.