Kibit Automator

RECENT UPDATES AND FUTURE DEVELOPMENTS

FRONTEO

KIBIT AUTOMATOR

- RECENT UPDATES AND FUTURE DEVELOPMENTS

Table of Contents

Summary

FRONTEO is a Japan-headquartered global technology organization that has been actively using and promoting artificial intelligence in business and legal operations for close to two decades. With its own development team and several data scientists on staff, FRONTEO has been actively engaged in developing and integrating AI-powered solutions in legal, compliance, healthcare, and business intelligence verticals.

Seeing a need for effective language analysis solutions for APAC clients due to the complexities presented by Japanese, Korean and Chinese (CJK) languages, FRONTEO has been working to custom-tailor its software to address complex language related nuances and focus on removing barriers to effective cross-border litigation.

FRONTEO's flagship product - an AI engine called "KIBIT" - has been in use by a multitude of large corporate and law firm clients since 2012 and has received multiple significant upgrades in recent years. KIBIT AI currently features several distinct AI algorithms with core capabilities in effective CJK language analysis. This paper will provide a quick overview of the major developments and updates in KIBIT AI over the past 2 years and will focus on its features and applications in the field of electronic discovery.

AI In Business – Uses, Metrics and Challenges

Overview of Potential Applications

Machine learning is a comparatively new technology, but it has convincingly shown its value time and time again in almost every area of life and has firmly entrenched itself as a mainstay business tool of the future. Machine learning and the AI-driven applications it enables, have a multitude of uses in addressing modern business needs - particularly in compliance, legal review, and business intelligence.

In compliance, AI can help organizations monitor their operations and identify potential regulatory issues by analyzing large volumes of data in real-time using up-to-date tools that ensure adherence to the latest regulations to avoid costly fines and penalties. It can also help rapidly complete annual data audits required by large government agencies.

In the legal field, AI can assist lawyers by analyzing document collections subject to litigation and identifying relevant content within documents, providing a major positive bottom-line impact by saving attorney time and reducing risk of errors.

Both law firms and corporate of-counsel can also leverage AI's ability to quickly structure unstructured data and visualize patterns to rapidly and efficiently complete internal investigations – especially when faced with regulatory pressure.

Finally, in business and economic intelligence, AI can analyze large datasets to identify trends and patterns, helping companies make informed decisions when it comes to supply chain management and stock ownership.

Overall, AI has the potential to revolutionize many aspects of modern business, and companies that embrace this technology are likely to gain a competitive advantage in the years to come.

FRONTEO, as a pioneer in the AI development market in Japan, has developed its own flexible AI engine called "KIBIT" in response to various needs. Although KIBIT's roots lie in the legal field, it has now been successfully adapted and used to deliver solutions for organizations that maintain practices in healthcare and life sciences, compliance, patent law, business intelligence and economic security. KIBIT's signature specialty is effective and accurate natural language processing and analysis of CJK language data. This paper will cover the developments in legal KIBIT AI (referred to as KIBIT Automator or KAM for short) specifically.

Measuring AI Efficiency and Accuracy

The decision to bring AI onboard your corporation's tech stack may not be a cheap one, and as with any other important acquisition it needs to be analyzed for risk and potential return. With AI, some of the metrics that decision-makers consider include precision, recall, processing speed, and AI training cost. High precision indicates that the AI system produces accurate results with a low rate of false positives. High recall suggests that the AI system can identify and retrieve most relevant information from a dataset. Processing speed affects how quickly an AI system can process large amounts of data; whereas training cost indicates the amount of money or time the corporation must spend on calibrating their AI system before they can begin realizing its full benefits. An ideal AI system would be one that requires minimal human supervision and can quickly learn and adapt to new data sources.

Good performance across all the aforementioned metrics is critical for businesses that want to get the most out of their AI systems. High precision and high recall ensure the AI system produces accurate and relevant results, leading to better decision-making and improved business outcomes. High speed is essential for businesses that need to process large amounts of data quickly, such as in cases involving real-time compliance monitoring or rapid regulatory investigations, and time is, frankly, always of the essence.

Cognizant of these priorities, FRONTEO has dedicated substantial development resources to continuously improve and iterate upon proprietary technology to create machine learning algorithms that deliver high-precision and high-recall results, while keeping training costs to a minimum. This paper will cover algorithmic performance improvements and provide further details in the sections below.

Language-Driven Complexities

The underlying performance of machine learning algorithms that drive a lot of the language-based analysis may be uneven across different languages. This should not be surprising – after all, syntax, grammar, and sentence structure can be quite varied even among related languages, not to mention something as dissimilar as Chinese and English.

In fact, interpreting CJK (Chinese, Japanese, and Korean) languages with AI can be particularly challenging due to their complex writing systems. These languages are characterized by lack of space separation, high character counts and flexibility in word ordering. As an example, while English builds its words and sentences from only 26 letters, Chinese has over 20,000 characters (although many of them might be infrequently used), and Japanese sees over 2,000 characters in daily use. Additionally, many of these characters will have multiple meanings based on their context and position in a sentence, adding further complexity to accurate meaning interpretation.

Another added challenge is that CJK languages have a unique grammar structure that differs significantly from English and other European languages. Chinese has no tense and no plural form of words. Both Chinese and Japanese have different word orders compared to English, and less rigid word order in a sentence. Korean has a complex system of particles that determines the grammatical function of a word in a sentence. These differences in syntax and grammar make it challenging to adapt AI models developed for Western markets to Asian data.

KIBIT AI – Answer to Handling CJK content

FRONTEO recognizes that developing accurate and effective AI systems for CJK languages requires extensive linguistic and cultural knowledge and poses unique challenges compared to languages with a more defined syntax structure. To address these challenges, FRONTEO has invested heavily in developing advanced AI technologies, including natural language processing (NLP) algorithms and machine learning models, that are specifically tailored to handle CJK data in addition to being fully competitive with other AI models on the market when it comes to handling English language.

Improving Japanese Parsing and Processing

Japanese language processing requires two core technologies – one is the decomposition of sentences into individual words and other morphemes (known as tokenization and morphological analysis) and the other is the analysis of the resultant morphemes to derive meaning.

The application of these technologies to Japanese is complicated by the nature of Japanese language as a non-segmented language – a language that does not utilize word breaks to delineate individual concepts or idioms within a sentence. Without those word breaks, which make tokenizing Western languages trivial, tokenization algorithms must be sophisticated, adapted to the specific linguistic content and context sensitive.

The accuracy of search and analysis is heavily influenced by the quality of tokenization. Poor tokenization can result in inaccurate search results, misinterpretation of text, and incorrect data processing. Consider an example – "会社員です" – the interpretation of this phrase would depend on whether it is tokenized as [会社, 員, です], where 会社 means "company" and 員 means "employee" or as [会, 社員, です], where 会 means "meeting" or "gathering," and 社員 means "employee" or "staff member." The correct answer would depend on the context and the intended meaning of the sentence. Each of these interpretations may result in a different treatment by AI in analyzing this concept and may affect search term results.

Furthermore, after the decomposition is complete, it is often difficult to evaluate the importance of single morphemes (such as "ha" and "ni") in the segmented text and properly weigh their impact on the overall relevance of a particular document.

With development of the "KIBIT" engine – FRONTEO's engineering team has leveraged over a decade of natural language processing experience to fine-tune our AI process to better handle tokenization output and expand CJK customized word dictionary to address its complexity directly and increase overall analysis accuracy.

In 2022, FRONTEO further released an update to its KIBIT algorithms to automatically identify and discard low-value single-letter words during the machine learning process, improving metrics like recall and precision on test data by approximately 7%.

On top of that, this update released in 2022 brought several other improvements to proprietary AI algorithms, not only in precision but also in computational speed – which has increased by a factor of 10 – enabling AI analysis of multi-million document sets to complete in a few hours at most.

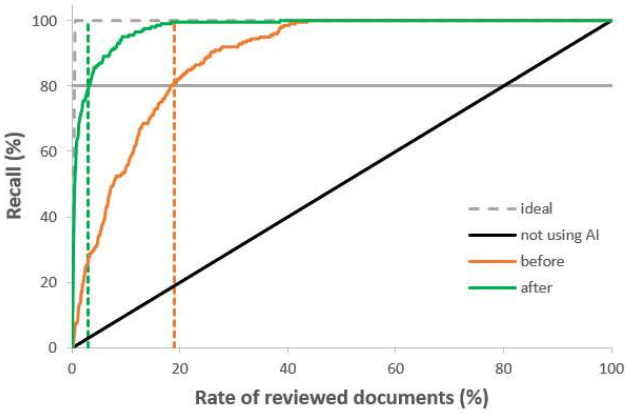

Figure 1: Improvements in AI precision over the last year

Figure 1 above shows the total improvement in precision of AI-driven analysis in legal document review through all of the improvements made to KIBIT in 2022. The graph indicates that to reach a level of 80% recall (which is equivalent to finding 80% of the relevant documents) with the improved algorithms, a human reviewer would be required to go through only approximately 3% of the entire data set vs approximately 18% with the older algorithms.

Algorithmic Innovation

FRONTEO's priorities with AI development remain focused on efficiency and accuracy of its "KIBIT" engine, and FRONTEO's sizable development staff and data science specialist are working on bringing a new algorithm to market in 2023.

The original algorithm behind "KIBIT" in 2012 relied on feature weighting to identify relevance in a data set in an effective manner with a small amount of training data and few positive training examples. The team innovated in 2019, developing a neural-network random forest-based algorithm to deliver extreme precision in datasets with medium to high richness, which had been a great success in delivering results to clients in a variety of cases.

FRONTEO is now bringing to market a new algorithm that combines the accuracy and low training data requirement of our previous algorithms. This new algorithm was inspired by intuitive cognition found in humans, and simulates the human ability to make snap judgments about color composition based on the hues of its individual components in a picture (see Figure 2)

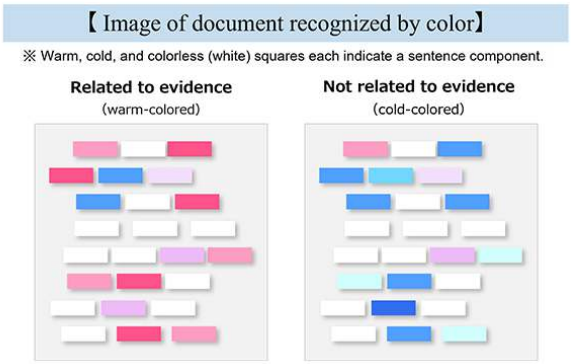

Figure 2: Each word is colored according to relevance, creating an overall feel for the document.

This algorithm will enable AI to score and rank the text of a document based on the rankings of individual morphemes found within those documents quickly and effectively. It also incorporates the various know-how and technological breakthroughs obtained in the development of the previous two algorithms used by KIBIT, combining the best of all worlds, and enhancing the features of each.

This new algorithm will be a general-purpose, high-performance AI engine that can serve as a singular go-to algorithm for providing fast and

accurate results in CJK languages. It will also feature a built-in display for confidence interval around the recall curve to improve defensibility of the AI review process.

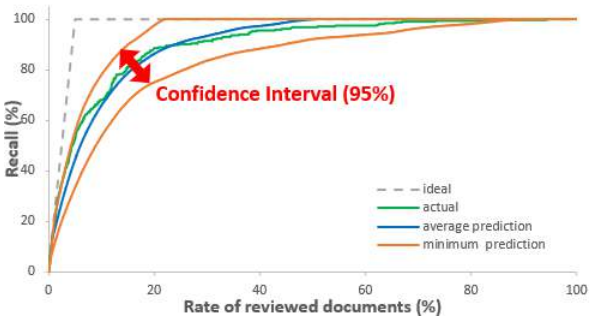

Figure 3: Confidence interval around a recall estimate in the new algorithm

Figure 3 above demonstrates the average recall curve for test cases utilizing the new algorithm (blue line) and how far it deviates from the actual results (green line). The lines are closely matched for most of the graph, indicating that the algorithm is extremely accurate at predicting the number of documents required for review before that review even takes place. This shows that KIBIT technology is capable of significantly alleviating the headaches of litigation logistics planning for document review projects by enabling legal operations teams to accurately estimate the extent of their document review burden before it takes place.

Automatic Sentence Highlighting

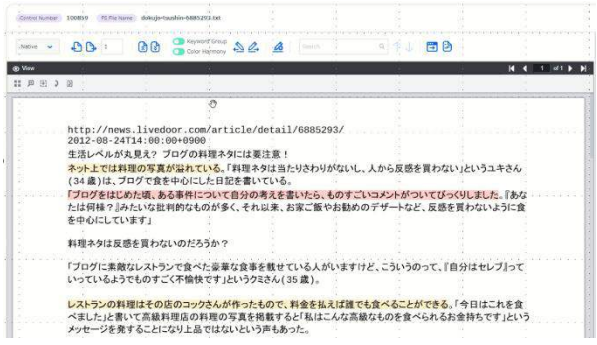

Figure 4. Example of AI-driven Highlighting in new algorithm.

The new algorithm not only improves on the document ranking process but can also score individual words within those documents and automatically highlight high-value sentences. Figure above shows an example. Sentences that are most relevant to a particular subject matter are highlighted in red, and those with moderate evidentiary value are highlighted in yellow.

Word Weights and Keyword Suggestions

To help with Early Case Assessment and Review phases of the EDRM model, FRONTEO is introducing several features in the immediate future that will help with identifying key concepts and words in document sets. Word weights show and visualize word distribution in the target data set, enabling attorneys to easily identify outlier words and concepts.

Keyword Suggestion will enhance the searching process by enabling KIBIT to automatically suggest additional keywords relevant to a particular query or a particular content type.

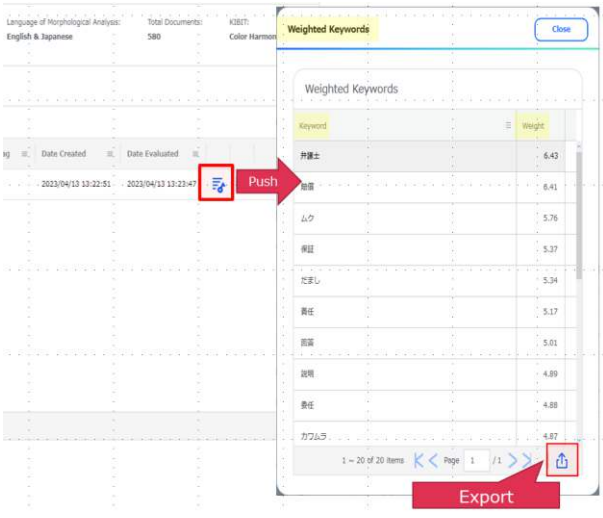

Figure 5: One of the features to be released in the next version of KAM that is designed to help identify key words and review documents

Data Assessment

A common challenge that our clients in the legal field often encounter are the so-called “rolling uploads” – situations where processing work for a particular forensic collection is done on an on-going basis. Their primary concern is whether the machine learning algorithms will have to undergo multiple training rounds to recalibrate them for additional documents. Previously, the answer would depend on client’s own estimates of how different the content or source of newly added data was compared to the existent sets. With a new feature that FRONTEO is introducing to KIBIT in 2023, the difference between data used to train AI and data that AI is being asked to evaluate can now be quantified numerically and KAM can produce an unequivocal answer as to whether additional sampling for training documents is required.

There are other potential use cases for this technique – for example some clients have inquired about building a persistent data library of training sets for different types of litigation or investigation content. This functionality would enable a quick assessment of how effectively new data can be trained using a pre-existing training set from the library.

Finally, FRONTEO is also working on utilizing this method for identifying the best or most optimal training set for a given document population, which would allow us to minimize the initial training process even further, driving further efficiency gains through the use of AI.

Conclusion

FRONTEO's flagship product – ”KIBIT” AI engine – is a prime example of the company's commitment to accurate CJK data analysis. KIBIT relies on advanced NLP techniques to effectively identify and categorize words and phrases in multiple languages, including Chinese, Japanese, and Korean. The engine features several machine learning algorithms that receive continuous support and updates over time to ensure they are always able to deliver accurate results with minimum time investment.

Moreover, FRONTEO's commitment to accurate AI analysis extends beyond its products and services. The company actively collaborates with academic institutions and industry organizations to advance research in AI and NLP technologies. This collaborative approach combined with FRONTEO’s investment in advanced AI technologies and a robust team of specialist developers and data scientists, all demonstrate the company's commitment to providing its customers with the most accurate and effective solutions to process, search and analyze CJK data.