MARCH 13, 2019

Legaltech® news

ONLINE

An ALM Publication

Natural Language Processing—Expanding Attorneys' Client Reach

With decades of natural language processing development have come some very simple yet powerful tools that can drastically cut the time and costs associated with complex legal projects.

BY PATRICK RYAN, STONETURN

The Chinese board game Go is one of the oldest games in the world. Enthusiasts have spent more than 2,500 years developing strategies to beat their opponents, and with hundreds of options for every move, its complexity can test the upper limits of human thought. So when Google's AlphaGo, an artificial intelligence (AI) program, defeated Go's world champion in March 2016, technologists took notice. Not only was the victory a testament to the power of AI, they argued, it was a bellwether of things to come, a demonstration that even highly skilled employees might soon be trampled in the unstoppable march toward an automated world.

The jury is still out on whether pure AI—that is, a technology that mimics the full complexity of human thought—is inevitable or mere hype, but earnest work is well underway and already paying off. In fact, technologies developed through AI research have been available for various aspects of legal projects for many years, though their relevance may not always be obvious.

Here, we explore a sub-field of AI, known as "natural language processing" or "NLP," to demonstrate how it can already make a difference in fact-intensive phases of complex legal projects.

Teaching Computers to Communicate Like Humans

NLP aims to provide computers with a human-level understanding of language. Researchers developed the first versions of NLP technology in the 1960s, but soon learned that teaching computers to communicate like humans was much more challenging than they had first thought.

Some difficulty lies in the nature and amount of information that people use to process language. If one were to say, "The Knicks really caught fire in the second half last night," a sports fan would quickly interpret that to mean the New York Knicks shot well in the second half of last night's basketball game. For computers, however, this is not so easy. Without information on the figurative meaning of the term "caught fire," for example, a computer might conclude that the Knicks literally went up in flames, even though such an event would be highly unlikely (even for the Knicks).

This complexity has led researchers to focus instead on solving more clearly defined problems. One such problem, commonly referred to as information extraction, is how to convert unstructured text into machine-readable, structured data. In corporate investigations, for example, a passage of text in employee expense data might describe the details of a particular business dinner. Information extraction algorithms can convert that passage of text into a relevant, structured record of information detailing who attended the dinner, the organizations to which those attendees belong, and where the dinner took place. In cases involving thousands of such expense entries, information extraction can prove invaluable.

Methods of Information Extraction and "Fuzzy Matching"

There are two primary information extraction methods used to determine such details: named-entity recognition and relation extraction.

Named-entity recognition identifies real-world objects referenced in a passage of text and classifies each of those objects as a particular type. For example, in an investigation into employee conduct, a named-entity recognition algorithm can help extract the named objects listed in the employee's expense details and classify them as persons, places, or organizations. Instead of a block of descriptive text, one can quickly see a list of attendees, organizations and establishments referenced in that text. For investigations involving multiple employees or thousands of expense entries, such an algorithm can help to extract valuable information in a fraction of the time that it would take for a team of trained investigators to do the same.

Relation extraction takes the information extraction process one step further by associating a list of names extracted from a passage of text to one another. A relation extraction algorithm first posits that a relationship might exist between every entity pair in a particular sentence. It then uses a combination of handwritten rules and statistical methods to determine whether each of those proposed

relationships actually exists, and if so, what type of relationship that pair shares. Extending the example of the expensed business dinner above, relation extraction can tell an investigator not only that the names of an individual and an organization appear in a passage of text, it can report to the investigator that the individual is likely an employee of the named organization.

Named-entity recognition and relation extraction are valuable tools for extracting entities and relationships from passages of text, but oftentimes relevant information becomes apparent only after connecting those named objects to external data sources. In an employee fraud case, for example, an employee's business expenses might reference a seemingly innocent meal with a business contact. However, if an investigator can determine the employer of the contact or related organizations, such as through the use of corporate registration data, the investigator might be able to gather valuable insights or make connections that would not otherwise have been apparent. Because names can vary in spelling and format from one source to another, such as an expense description versus a corporate registry, it can be tedious and time consuming to match names manually. To help automate this process, a data analytics expert can use an approach commonly referred to as "fuzzy matching."

"Fuzzy matching" is an NLP process that finds approximate, as opposed to exact, text matches to connect names and relationships to information found in other documents or text. It allows an analyst to determine the likelihood of two text strings being the same, even if one of the text strings is misspelled or slightly different than the other. Using a fuzzy matching algorithm, an investigator can identify a possible match between two names even if those names are not quite the same. The algorithm can even provide a metric for how close the two names are to each other, allowing an investigator to eliminate false positives before they are manually reviewed in a verification process. This is extremely valuable when dealing with complex matters involving dozens or hundreds of actors or entities.

NLP is a fast-developing area of AI, with exciting new applications and algorithms appearing every week, but it is important to note that NLP has decades of development already under its belt. With those decades have come some very simple yet powerful tools that can drastically cut the time and costs associated with complex legal projects.

Patrick Ryan *is a managing director with global advisory firm StoneTurn. Based in New York, he provides data science and analytics services to clients in the legal and financial services industries.*