Research Article

Advances in Machine Learning & Artificial Intelligence

Semantic NLP Technologies in Information Retrieval Systems for Legal Research

Sarvajna Kalva* and Fred Geldon

Computer Science Department, George Mason University, USA

***Corresponding author**
Sarvajna Kalva, Computer Science Department, George Mason University, USA

Submitted: 17 July 2021; **Accepted**: 28 July 2021; **Published**: 05 Aug 2021

Citation: *Sarvajna Kalva and Fred Geldon (2021) Semantic NLP Technologies in Information Retrieval Systems for Legal Research. Adv Mach Lear Art Inte, 2(1): 28-32.*

Abstract

Companies involved in providing legal research services to lawyers, such as LexisNexis or Westlaw, have rapidly incorporated natural language processing (NLP) into their database systems to deal with the massive amounts of legal texts contained within them. These NLP techniques, which perform analysis on natural language texts by taking advantage of methods developed in the fields of computational linguistics and artificial intelligence, have potential applications ranging from text summarization all the way to the prediction of court judgments. However, a potential concern with the use of this technology is that professionals will come to depend on systems, over which they have little control or understanding, as a source of knowledge. While recent strides in AI and deep learning have led to increased effectiveness in NLP techniques, the decision-making processes of these algorithms have progressively become less intuitive for humans to understand. Concerns about the interpretability of patented legal services such as LexisNexis are more pertinent than ever. The following survey conducted for current NLP techniques shows that one potential avenue to make algorithms in NLP more explainable is to incorporate symbol-based methods that take advantage of knowledge models generated for specific domains. An example of this can be seen in NLP techniques developed to facilitate the retrieval of inventive information from patent applications.

NLP, AI, and the Shift Towards Machine Learning

Natural language processing (NLP) is the application of techniques that enable computers to interact with human language. Artificial intelligence and NLP are often mentioned together; in fact, NLP can be seen as a branch of AI, because it helps achieve one of the goals of artificial intelligence. If the goal of AI research is to develop computing machines that think the way humans do, the goal of NLP research is to develop computing machines that can understand and generate language the way humans do.

Information retrieval systems are used by lawyers to access particular case decisions, look up definitions or legal standards, and generally figure out what precedent has been set to date on a given legal issue. The listed examples have clear and specific applications to the kind of research conducted within the legal field. but developing applications to meet the general goals of legal research can be tricky. An Above the Law article quotes Justice Felix Frankfurter, who once stated, "Research requires the poetic quality of the imagination that sees significance and relation where others are indifferent or find unrelatedness; the synthetic quality of fusing items theretofore in isolation; above all the prophetic quality of piercing the future by knowing what questions to put and what direction to give inquiry.". While these underlying goals inform research and analysis for the legal profession, Frankfurter's aspirations for a legal researcher can also be applied as the ideal goals for technology used to aid legal research.

A well-functioning information retrieval system will be able to identify relationships between the information relayed in different documents, put the disparate connections in context with one another, and be able to predict the questions one would have about a given topic. The transformative ability of computers to store and perform computations on large amounts of data has led to sophisticated information retrieval in computerized databases of legal documents becoming a mainstay in the legal world. However, evaluating them by Frankfurter's research guidelines requires a development of semantic context for language and an understanding of the meaning within a text, and our computer algorithms are not yet powerful enough to fulfill these goals to the extent that humans can. Still, the listing of potential NLP applications by the Department of Computer Science and Technology at Tsinghua University shows an optimism for what they term "LegalAI" (legal artificial intelligence); the enumerated possibilities, some of which are no longer in the beta stages of development, range from summarizing text in legal documents and matching similar cases all the way to AI that can answer legal questions and predict court judgments based on settled precedent [1]. The aim of this survey will be to explore how NLP has played a role in helping information retrieval systems become more intelligent, and what else can be done to improve their functionality for use in legal research.

Types of NLP Techniques

Lupu et al. describes the evolution of NLP tools as one that began

from hand-coded linguistic rules [2]. These rules were replaced by machine learning algorithms as the increased computing power of machines enabled the handling of the large sets of training data required for learning. Machine learning is the process through which a computer gets better at guessing the correct answer for a given problem by making many, many guesses to questions/answer examples and making tweaks to its prediction model along the way.

Like the development of computerized databases for legal research and the large-scale application of machine learning, the use of NLP in law is a relatively recent phenomenon. The adoption of the NLP technology has been led by companies like Westlaw, Lexis, Bloomberg, Fastcase, Ravel, and Case text, which have the substantial advantage of already possessing a large corpus of annotated legal text. While these legal information retrieval systems are proprietary, Callister makes some informed guesses at how NLP is incorporated within legal research systems such as Westlaw and Lexis Nexis by pointing to the kind of analysis that is conducted on legal documents. For instance, he states that "certain natural language processing activities, such as classification, are useful in determining negative and positive citations to a work". Callister is referring to the sherardization feature in LexisNexis, which provides guidance as to whether the precedent set by a case is still valid, based on the treatment it was given and the decisions that were ultimately made in other cases that cited the precedent. This feature most likely uses sentiment analysis, an NLP technique that characterizes subjective information within language by classifying text into categories like positive, negative, or neutral.

Today, most modern NLP tools involve some form of machine learning. NLP tools can be developed to be used generally as an off-the-shelf product or developed for a specific company or domain. The development choice that is made will impact the accuracy and applicability of the tool for its intended purposes. General NLP tools are often limited in the depth of their linguistic analysis, because working with lots of unseen semantic context can lead to a decrease in performance, "since it involves manual annotation of training data and creation of a ground truth for evaluation" [2]. Training data for NLP is labelled with linguistic annotations, metadata that provide additional information for identifying the patterns a model is supposed to learn. This annotated data is useful for training, because legal documents are generally stored in the form of unstructured text. The language is encoded as symbols without any meaning. Examples of linguistic annotations may be syntactic, such as the grammatical tagging of parts-of-speech, or semantic, such as sense-tagging, the assigning of lexical categories to words. Linguistic annotations can also be used for evaluation of machine learning models. Sentiment tagging a text, for instance, allows models conducting sentiment analysis to test the accuracy of their predictions about the sentiment in the language. There is limited availability in the quantity and diversity of large databases of text that is linguistically annotated within the legal field, because linguistic annotations are often manually added by humans. The resources involved in making these annotations is another roadblock that must be dealt with for NLP. Without the appropriate databases for training and evaluation, the resulting models would have less accuracy.

Knowledge Models vs Big Data Methods

In the aforementioned methods, NLP analyzes the formal language within legal texts in order to extract meaningful information. But although the need to semantically par text to gain a deeper understanding of language theoretically makes sense, the details become much less straightforward in implementation. Lupus et al use Wittgenstein to describe the fundamental basis of some kinds of semantic technologies, stating that they are "essentially grounded in Wittgenstein's observation that the meaning is defined by usage" [2]. The reference to Wittgenstein can also describe the essential motivation to apply semantic technology to search systems. Taking advantage of semantic context allows algorithms to better represent significance in the meaning of words, because meaning in language arises through context, the ways in which words are used.

While there were many kinds of semantic NLP techniques mentioned by the authors included this survey, who often use different terminology to describe their approaches, two distinct categories of analysis emerged throughout. Whether the different approaches to NLP technology were termed knowledge-oriented vs data-oriented, statistical vs explicit, or symbol-based vs embedded, they all distinguished between NLP that depends on knowledge models as opposed to statistical inferences made using massive sources of data for its functionality.

Lupu et al. outline two kinds of "semantic" technologies: statistical and explicit. Semantic context can be used to describe metrics such the frequency of words occurring in text data, or it can be used to refer to knowledge bases built manually on top of the data from a specific domain [2]. In explicit semantic technologies, a knowledge base for the semantics of the data is developed before further text analysis. The goal of the knowledge base is to make explicit the concepts and relationships agreed upon within a domain; they can be as simple as a glossary of terms and definitions or be designed to represent more complicated ontologies/taxonomies, where items are linked through class hierarchies and other relationships. The knowledge bases are built on top of unlabeled and unstructured text data. One benefit of explicit approaches is that knowledge bases give the conducted analysis more visibility. While they are still often combined with models which make decisions that are not intuitive to users, those involved in the backend development of software systems can exchange and make use of the information contained in the knowledge bases to understand what assumptions have been made within the algorithm. The assumed entities and relationships that have been used to make sense of the data are more transparent to the user.

Zhong et al., who term the application of AI to legal tasks as "Legal AI", divide their discussion of NLP methods into the categories of symbol-based methods and embedded-based methods [1]. Symbol-based methods take advantage of symbols from the legal domain to create a basis of legal knowledge for application within LegalAI. An example of symbol-based methods is legal element extraction, which takes text as input and extracts information pertinent to specific legal elements. For instance, in a case about fraud, information about the elements that constitute fraud might be extracted from a text containing the facts of case: whether there

was a representation, a false representation, or knowledge and intent by speaker of said false representation, among other elements. More general kinds of information may also be extracted from a text with symbol-based methods, such as relations between people or timelines for factual events [1]. Because legal concepts can change so much between different countries, the most successful information extraction models operate on specific domains.

Embedding-based methods, on the other hand, represent legal facts and knowledge in "embedding space" and can use deep learning to improve the completion of tasks. Like Callister, the computer scientists acknowledge that embedding-based methods "bridge the gap between texts and vectors", however, they see it as a positive feature that allows for more sophisticated algorithms [1]. While embedding-based methods have better performance, they lack interpretability. The results will be determined by models which make tiny adjustments until the correct predictions are given. The prediction will eventually be correct, but the process leaves no line of reasoning to help the user understand how the determination was made.

The Transparency/Efficiency Tradeoff

The next shift in NLP is predicted by Lupus et al. to be in the area of statistical semantics, a data-oriented approach including methods like unsupervised deep learning [2]. Deep learning is a more sophisticated kind of machine learning, in which the model is able to not only change the predictions it makes, but also intelligently modify the process by which it learns, based on observed patterns during training. While deep learning has brought tremendous progress to the field of AI in the recent decade, there also exists a tradeoff of having algorithms that are explainable in a way that can be interpreted by humans. Paul Callister, a professor of law, accepts the influx of AI in the legal profession, but is concerned with the idea that lawyers will come to depend on a source of knowledge with which they have little control or understanding. In his paper, Callister focuses on how lawyers can learn to interact with NLP technologies through user training for accessing online legal research systems and an improved understanding of the technology that powers the systems. However, there is only so much that can be done from the user's perspective about the way that statistical semantic technologies operate using black box mechanisms, because a machine learning program obscures what assumptions were made by the model, which causes incorrect results to appear.

One example of a deep learning application is in latent semantic analysis, which builds up semantic meaning by comparing how frequently certain words appear together in a document. In an experiment on the use of latent semantic analysis (LSA) for text summarization, the authors Merchant & Pande create summaries that convey the significant concepts in a document. Like other LSA techniques, their model takes advantage of the assumption that "semantically similar terms will occur in similar pieces of text" [3]. The words within a document are represented by a set of concepts, or word phrases. What groups of words will appear together as concepts will not be known before the analysis, but the clustering of these groups of words develops as a result of an unsupervised machine learning process.

The categories created in the LSA do not necessarily correspond to any intuitions one would have, because the patterns that are computed cannot be predicted before the model is run. In supervised machine learning, where the model is given a set of categories and then trained to classify inputted data into the proper categories. However, when the categories are not known, they have to be created by the model as part of the learning process. In unsupervised machine learning, the training of the model happens without any pre-assigned labels that could be used to categorize the data. The categories have to be created on the examples the model sees during training. Furthermore, like other machine learning models, the kinds of statistical decisions made in LSA that lead to the results are not accessible to a programmer. The model developed in Merchant & Pande eventually produced summaries such as "The Trial Court found all the accused guilty of the charges and convicted and sentenced them" [4]. If any of the summaries were false or misrepresented the text, it would not be clear what caused the incorrect summary to be produced by the model. The inaccuracy might be the result of bias in the NLP model or in the data that the model has trained on, but unless the model programmed to explain its conclusion, the patterns found in machine learning are represented in statistical calculations opaque to the user.

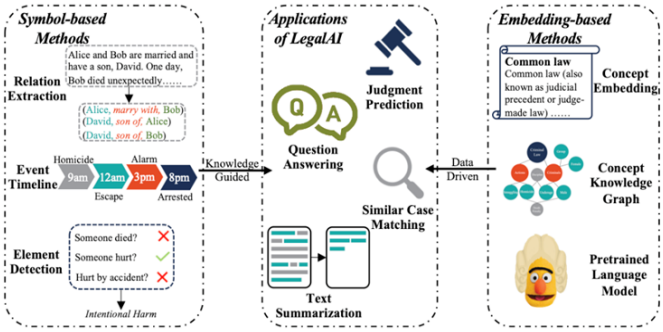

Figure 1: An overview of tasks in LegalAI.

Note: A breakdown of types of NLP techniques. Reprinted from "How does NLP benefit legal system: A summary of legal artificial intelligence." by Zhong et al, 2020 [1].

However, the key to navigating the tradeoff between the inefficiency of knowledge-oriented approaches and the lack of interpretability in data-oriented approaches is to incorporate both in NLP techniques. Zhong et al. believe knowledge modelling may improve interpretability for embedded approaches in applications like the detection of legal elements in text, because it is "impossible to learn the meaning of a professional term directly from some legal factual description" [1, 5]. There must be another layer which adds context to the factual descriptions. Knowledge graph methods are one example of an approach that can fill this gap, taking legal concepts and encoding their meaning by making connections to other entities they are related to. As a rudimentary example, the concept of 'criminal law' might be linked to 'actions' like 'smuggling' and 'homicide' and the concept of 'homicide' might be linked to 'death penalty' (Figure 1). Knowledge graphs should be tailored to break down legal concepts so they can map better to legal text. In symbol-based methods, the extraction of symbols like legal elements, events and relationships from legal

document provide more interpretability. Both embedded and symbol-based methods must be incorporated into AI techniques if they are to be applied in real-world legal systems.

NLP Applications in Information Retrieval Systems for Patent Law

Deep linguistic analyses work most optimally in specific domains, where the targeted nature of the training data leads to a better ground truth on which to evaluate a model's predictions. The application of NLP to the legal domain marks one level of specificity for the type of data models train on, but optimally the domain of application might be restricted to a particular legal branch like intellectual property, or focus even further on patent documents, a specific type of IP. Applying NLP techniques in patent corpora can still be particularly interesting because patent applications are required to follow certain legal guidelines and stylistic formalities in order to be accepted. This makes the language somewhat more predictable than judges' opinions or other such legal documents.

Patent texts are still considered to be unstructured text, however, since the words in patent applications do not come linguistically pre-annotated with known entities and concepts. In the paper, "Improvement of Automatic Extraction of Inventive Information with Patent Claims Structure Recognition", Berduygina & Cavallucci incorporate a "knowledge-oriented approach" into their model, which can be more useful when there is unstructured data that lacks uniformity. Different inventors in patents often use different terminology, even when they exist within the same field of art or when describing similar inventions, so patents lack lexical uniformity. These considerations shape the type of NLP techniques used by the authors in their application. The NLP behind their extraction process clearly takes advantage of formal linguistic patterns within the patent text, but these patterns have to be detected from scratch. The knowledge-oriented approach played a role in detecting existing linguistic patterns and adding a contextual layer of meaning to the unstructured data.

Berduygina & Cavallucci look at patents specifically because "The formulaic language used in patent claims construction enables [us] to say [there are] determined dependency constructions". Patent texts are unique in that they must meet certain requirements regarding the layout of the content. A typical patent document contains series of claims at the end of the document including descriptions such as the category, purpose, and technical features of the invention. The claims depend on each other, and language like "Claim 2: ...as set forth in Claim 1..." makes these dependencies explicit. The formal language used in the claim construction that the authors are referring to lends clarity to the way the different ideas within the description of the invention relate to each other. The authors propose that understanding the full hierarchical structure within the way the claims depend on each other can improve the time and accuracy of the information extraction.

Once the hierarchical structure of the claims was analyzed for their dependencies on one another, the inventive information still has to be extracted. In order to filter the general information contained in the claims to pull out the inventive information that is desired, the authors used a model to automatically extract three components of the inventive information: describing the problem which the inventor is trying to solve or describing part of the solution the inventor has come up with. The patterns for the problem match the following linguistic structure: <subject> doing some <action verb> in a <complement, ie. noun or adjective> explaining the situation. For instance, the entities within the sentence might be: <corn chip fryers> <fail> to <fry chips in a uniform, stackable manner>. This linguistic pattern is a generalization of the common grammatical structures in "an obstacle prevents progress" the phrase Berduygina & Cavallucci identify as the problem in the patent application [3]. A solution takes the form <verb> + <complement>, as in "expresses a result". This may look like something like: the claimed invention <cooks the chips using two mold cavities to constrain said chips>. The extraction of these problem and solution patterns (which could be composed of words or contiguous word phrases) is based on the frequency in the occurrence of the markers within the particular document as compared generally to other documents in the patent corpora.

Berduygina & Cavallucci's experiment combines portions of a data-oriented approach with the knowledge-oriented approach [3]. The automatic extraction looks at elements like word frequency, similar to the data-oriented latent semantic analysis method, yet the extraction also prioritizes the problem-solution linguistic patterns that occurred more often in the document by assigning them more weight. The authors note that their method improved the model's extraction by reducing the time required to process the claims and quality of the solutions that were captured from the claims. However, there was noise; the information extraction yielded better results, but not all of them were relevant to inventive information they intended to extract. This was probably due to the small size of the dataset, which caused the model to be less precise. The final API was designed to take a patent text as an input and extract inventive information from that text based on the list of linguistic patterns that have already been extracted.

Patents carry descriptions of inventions, and inventors interested in patents must go through the process of researching what other patents exist before they seek to secure the benefits of their ideas for themselves. The authors chose this problem of mining for current inventions within the patent corpora because they believe tapping into this information would help people access the information for future research and spur on innovation. While the application of Berduygina & Cavallucci's paper is not strictly targeted towards the legal field, retrieving such information could be helpful to lawyers drafting patent applications. The tool can also help examiners who have to search for prior art, the process of checking what parts of an invention are novel and nonobvious to those of ordinary skill in the art and can be patented. The linguistic patterns used to extract the problem and solution represent a simple format the inventive information might take within a patent and show how a knowledge-oriented approach can improve the quality of information extraction. Furthermore, the linguistic patterns make explicit important assumptions made by the automatic extraction algorithm about what counts as "inventive information". In other legal fields as well, incorporating knowledge-oriented approaches to NLP into data-oriented ones in this way can improve the interpretability of NLP algorithms used in information retrieval systems and encourage the adoption of NLP technology for legal research [6-10].

References

1. Zhong H, Xiao C, Tu C, Zhang T, Liu Z, et al. (2020) How does NLP benefit legal system: A summary of legal artificial intelligence. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics 2020: 5218-5230.
2. Lupu M, Mayer K, Kando N, Trippe A J (2017) Current challenges in patent information retrieval (2nd ed.). Springer.
3. Berduygina D, Cavallucci D (2020) Improvement of automatic extraction of inventive information with patent claim's structure recognition. Advances in Intelligent Systems and Computing 2020: 625-637.
4. Callister PD (2020) Law, artificial intelligence, and natural language processing: A funny thing happened on the way to my search results. Law Library Journal 112: 161-212.
5. Ashley KD, Walker VR (2013) Toward constructing evidence-based legal arguments using legal decision documents and machine learning. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law – ICAIL 13: 176-180.
6. Ambrogi, B (2019) Legal analytics products deliver widely divergent results, study shows. LawSites. https://www.lawsitesblog.com/2019/11/legal-analytics-products-deliver-widely-divergent-results-study-shows.html
7. Black N (2019) Lawyers have a bevy of advanced and AI-enhanced legal research tools at their fingertips. ABA Journal. https://www.abajournal.com/web/article/lawyers-have-a-bevy-of-advanced-and-ai-enhanced-legal-research-tools-at-their-fingertips
8. Lat D (2018) How artificial intelligence is transforming legal research. Above the Law. https://abovethelaw.com/law2020/how-artificial-intelligence-is-transforming-legal-research/
9. Merchant K, Pande Y (2018) NLP based latent semantic analysis for legal text summarization. 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2018: 1803-1807.
10. Stede M, Schneider J (2018) Argumentation mining. Morgan & Claypool Publishers.