How Could Semantic Processing and Other NLP Tools Improve Online Legal Databases?

Renátó Vági

MONTANA Knowledge Management Ltd.
Budapest H-1097, Hungary

Doctoral School of Law
Eötvös Loránd University
Budapest H-1053, Hungary
vagi.renato@montana.hu

Abstract: The spread of online databases and the increasingly sophisticated search solutions in the past 10–15 years have opened up many new opportunities for lawyers to find relevant documents. However, it is still a common problem that the various legal databases and legal search engines face an information crisis. Legal database providers use various information extraction solutions, especially named entity recognition (NER), to mitigate this problem. These solutions can improve the relevance of the lists of results. Their limitation, however, is that they can only extract and create searchable metadata entities if the latter have a well-defined location or regularity in the text. Therefore, the next era of search support for legal databases is semantic processing. Semantic processing solutions are fundamentally different from information extraction and NER because they do not only extract and make visible and/or searchable the specific information element contained in the text but allow for the analytical analysis of the text as a whole. In addition, in many cases, legal database developments using machine learning can be a significant burden on a company, as it is not always known what kind of an AI solution is needed, and how the providers could compare the different solutions. Legal database providers need to customize processing their documents and texts in the most optimal way possible, considering all their legal, linguistic, statistical, or other characteristics. This is where text processing pipelines can help. So, the article reviews the two main natural language processing (NLP) solutions which can help legal database providers to increase the
value of legal data within legal databases. The article then shows the importance of text-processing pipelines and frameworks in the era of digitized documents and presents the digital-twin-distiller.

**Keywords:** digital-twin-distiller, legal databases, legal research, named entity recognition, semantic processing, text-processing pipelines

1. Introduction

Various technological tools and software that have the ability to facilitate legal work have undergone tremendous development in recent decades. The same applies to legal databases and legal search engines. In the last 10–15 years, online legal databases have become globally widespread following paper and CD-based or other optical media-based storage formats (Margolis & Murray, 2012). This means that online databases have become the primary source of legal information, and more and more old documents are being digitized and made available online.

At first, this seems to be a positive change, compared to previous solutions, as it may increase the availability of legal texts: documents are available in a few clicks, in many cases free of charge, and people can use them widely, even beyond the legal profession. It also means that lawyers do not need to go to a library or buy expensive software to reach what they are searching for, they can simply go online and find every statute, every court decision, and every other publicly available legal document.

However, in practice, we can see that the emergence of online legal databases contributes to a significant increase in legal data, which leads to the problem of the so-called information crisis (Bing, 2010). Information crisis, or information inflation, describes the process in which the amount of data in a given area suddenly increases, making it difficult to discover the data in its current form and requiring to switch to a new type of storage or retrieval method. However, this new type of form, in addition to making relevant content more transparent and searchable, also increases the speed and volume of data generation as an unintended effect, creating the need for more efficient search methods. Following the emergence of electronically stored legal information (Webb, 2020), this can also be seen during the emergence of online legal databases.
As a result, legal research could be a time-consuming and labour-intensive task. The main reason for this is that different legal databases and their result lists are over-inclusive and under-inclusive, especially those that use only keyword search. This means that it is a reoccurring problem that the result list does not contain all the documents that may be relevant but also contains many documents that are not relevant and only create noise for the legal research. The third problem is homogeneity, which means that the result list is not very informative because it mostly contains the number of the court decision or the names of the litigants, whereas that does not provide much information about the content or the relevance of the decision for the legal researcher (Walters & Asjes, 2021, pp. 358–359).

This is why there is a need for new kinds of text-processing solutions in the field of legal databases and legal search engines. To diminish these problems, legal database providers use various natural language processing (NLP) tools. Next to machine learning and expert systems, NLP is one of the major artificial intelligence subfields in legal technology literature. Artificial intelligence in legaltech is a broad field of software development, in which the goal is to automate a segment of the lawyers’ work process using computing capacity and the computers’ learning ability. Within this term, NLP is a set of tools that work with the linguistic aspects of a text. They can be used for information retrieval, information extraction, automatic summarization, sentiment analysis, etc. Within NLP, we can distinguish between solutions that use syntactic or semantic elements of the text, or a combination of the two (Katz, 2021). In a simplified way, the former can be called named entity recognition, and the latter semantic processing. The exact meaning of these two will be explained below.

The use of NLP technologies is a recent phenomenon, and mostly British and American legal data providers use it, as they have the required large corpora of legal texts (Kalva & Geldon, 2021). One of the main reasons for this is that British and American legal practice relies much more on case-law than on continental law, in which statutes and regulations are the primary legal sources. As a result, there are a lot more unstructured text files in the former, which makes it necessary to make their searchability effective because of the need to find them and know their legal content.

In continental legal databases, named entity recognition (NER) tools are usually implemented. These solutions can improve the relevance of the result lists, but their limitation is that they can only extract and make searchable metadata entities in the text. Therefore, the next era of search support for
these legal databases is semantic processing. My main research questions in this article are what is new about semantic processing compared to named entity recognition, how it can increase the efficiency of legal work, and how can different text-processing pipelines combine these solutions.

The article reviews the two main natural language processing solutions which can help legal database providers to increase the value of legal data within legal databases. It is organized as follows. Section 2 is about information extraction and named entity recognition, which are perhaps the more common methods. The article shows what they are exactly, how we can use them, and their main limitations. Section 3 reviews semantic processing, which is a relatively new phenomenon in legal technology, demonstrating its solutions and how it can transform the functions of legal databases soon, as well as how it can reshape the legal research and the work of legal databases editors. Section 4 is about the importance of text-processing pipelines in the era of digitized documents, and the article presents the digital-twin-distiller, an open-source framework that can integrate the positives of NER-based and semantic-based solutions.

2. Named entity recognition-based solutions for legal databases

Legal research is one of the most versatile and time-consuming parts of legal work. According to the 2019 Bloomberg Law survey, the participants considered it the fifth most time-consuming litigation task. Respondents said that the main reason for this is that it is difficult to tell where legal research ends and where the following work process begins. In addition, it is often necessary to return to it because online legal information grows in volume and complexity, and lawyers cannot afford to miss some significant changes (Bloomberg Law, 2020).

In addition, this term can be used in two meanings. On the one hand, it represents the well-defined part of the workflow of legal practitioners when gathering the legal provisions relevant for their current case. It can happen at any stage of a given case, even if they provide only simple advice to their client, or only draw up a contract. However, it plays a significant role when they prepare for a lawsuit and seek legal arguments for why the court should decide in favour of their client (Heller & Arredondo, 2021). Due to the characteristics of the continental legal culture, this mostly means looking
for legislation. However, searches for court decisions, judgments of other law enforcement bodies, and scientific legal literature are also important.

On the other hand, legal research also refers to academic legal research, which means not so much the search for relevant case-law, but rather the statistical analysis of the decisions taken by legislation and law enforcement agencies, and the creation of policy proposals (Custers, 2018, p. 2). We can also describe this by the terms of the sociology of law or empirical legal research. Of course, these two meanings are not strictly separated. Furthermore, as the statistical or data-driven approach to legal texts begins to gain ground among legal practitioners, more and more empirical and in-depth analytical tools are becoming widespread. In addition, both rely heavily on online legal databases, so that the operation of these databases largely determines the effectiveness of both kinds of legal research.

Continental legal database providers, if they use any NLP technologies, mostly use information extraction and NER solutions to increase the efficiency of their legal search engine. This section presents what these solutions are, what they can be used for, and what their limitations are.

Information extraction means finding and extracting specific elements of information from texts. The essence of these solutions is that if we know the regularities of a given type of information within the text, they can be automatically extracted from the text and then made available for any later use (Sakhaee & Wilson, 2020). Named entity recognition is one of the main areas of information extraction, and one of the most used fields of application of natural language processing. In this case, the information content to be recognized and extracted is a named entity. In the case of legal texts, it is usually a reference to the law, a reference to a court judgment, the date, the names of the litigants, or their representatives and judges, or even definable legal terms (Nadeau & Sekine, 2007).

As far as legal databases are concerned, the most obvious way to use these solutions is to make it possible for legal information entities to be retrieved automatically, without editorial help or any other intervention, and then make it searchable by the users of the database (Iftikhar, Ul Qounain Jaffry & Malik, 2019; Sharafat, Nasar & Jaffry, 2019). It gives the user more filtering options than a keyword search, making it easier to find the legal document that is relevant to them. So it can reduce the over-inclusivity and the under-inclusivity of the result list. It can also facilitate the work of the editors and make the processing of legal data faster.
Another way to use NER for legal text is creating citation networks for legal texts. This means that a reference in a legal document to another legal document (such as case-law or statutes) can be retrieved automatically. These could then be used to create a network of legal documents. On the one hand, this links legal documents to each other much more visibly. So, if users find a document relevant to their case, they do not have to make a new search to find similar documents but can switch to other documents related to the original document with one click. On the other hand, creating a citation network between legal documents opens up new possibilities in legal document analysis, thus fundamentally transforming legal research (Olsen & Küçüksu, 2017) because legal researchers can analyze the system of legal regulation and case-law and the contradictions between them far better, and they can find the legal document that contains the most relevant regulation or argument for them much more quickly and easily. In addition, legal databases can use the extracted citation network to determine which statutes and judgments have been cited the most, and the system can prioritize them in the result list (Doslu & Bingol, 2016; Ashley, 2019). Thus, the user can see the most relevant documents in the result list, which reduces the research time.

Another interesting solution between the areas of named entity recognition and semantic processing is semantic metadata extraction. This means that not only the named entity element itself can be extracted from the text but also the role of the word or phrase played in its semantic meaning (Sleimi et al., 2021). Thus, it can be determined, among other things, whether a particular piece of legislation contains an obligation, a right, a prohibition, or a permit, or whether the extracted data in the legal text refers to an actor, place, action, date, etc. applies. It will allow for deeper automated analysis of legal texts and more customizable legal databases that allow relevant legal information about a particular well-defined clientele to be included in one place (Zeni et al., 2013).

Thus, named entity recognition solutions can be used to increase the efficiency of legal databases. However, a major limitation of these solutions is that they can only be used for explicit content in texts, and only if they have a well-defined location or regularity. Fortunately, there is another area of use of natural language processing for legal texts, which is the so-called semantic processing. The next section presents what this phrase exactly means and how it can reshape legal research and the work of editors of legal databases.
3. Semantic processing of legal text

Semantic processing differs from information extraction and named entity recognition in the way that it does not only extract elements of information contained in the text explicitly and with well-defined regularity, but it also allows a deep analysis of the legal text as a whole (Francesconi et al., 2010). In short, it helps the algorithms of the legal databases to represent the meaning of the legal text better, even through context (Kalva & Geldon, 2021).

In addition, the automatic semantic processing of legal texts paves the way for a deeper semantic analysis of the law, providing new opportunities both for a deeper understanding of the legal domain and for developing a more appropriate litigation strategy. Tools such as TF-IDF, LSI, and LDA can help achieve this. Term frequency inverse document frequency (TF-IDF) is a simple, long-known method of vectorization that statistically weights words in a document based on how often the word occurs in a given document (term frequency) and the overall data set (inverse document frequency) (Robertson, 2004). The downside, however, is that this method is only able to examine individual terms on a statistical basis and is unsuitable for mapping the semantic relationships between words. Latent semantic indexing (LSI), on the other hand, can complement this method by making the coexistence of terms statistically detectable, making their semantic context more mappable. LSI is also a good tool for recognizing synonyms and polysemy words (Deerwester et al., 1990). Finally, latent Dirichlet allocation (LDA) is a topical modeling method, based on a three-level hierarchical Bayesian model, that can specify the approximate number of topics of a dataset and the keywords associated with each topic. It can be used on a single document if we want to know the main topics in a text, or on multiple documents to check how many different groups they contain. Thus, documents that are close together in their semantic meaning, are much easier to discover (Blei, Ng & Jordan, 2003).

The applicability of this type of solutions in the legal field is primarily the automatic categorization of different documents (Francesconi & Peruginelli, 2008; Orosz et al., 2021). It allows various legal documents to be grouped not only according to the regularities of current legal dogmatics but also according to the use of other aspects, for example, statistical regularities. In addition, it also makes the searches interoperable between different sources of law. In legal databases, different legal sources, such as legislation,
court decisions, scientific legal literature, etc., are treated separately for the better transparency of the documents. With semantic legal document categorization, however, the legal sources could be searched in an integrated way.

Another fruitful area of semantic processing is automatic summary generation (Chieze, Farzindar & Lapalme, 2010; Kanapala, Jannu & Pamula, 2019). As mentioned above, one of the shortcomings of legal databases and their result lists is the so-called homogeneity, so the result list does not contain much information about each document, and thus it is difficult to decide its relevance to a given user. In the case of court decisions, a summary of the case with all the essential elements can help this so that it will be possible to decide in a short time whether the document is relevant to us or not.

Finally, semantic processing is perfectly suited to support similarity searches and text similarity analysis between legal documents. Generic search engines like Google have supported features that allow users to search for similar documents with results that have been relevant to them for a long time. However, most legal search engines do not currently support this. A significant part of the legal work is finding similar documents to their case or another document, so automating this task could decrease the work of lawyers. For example, a significant part of legal research is looking for court decisions that could be relevant to a client’s case. Similarity recommendations can help lawyers find similar documents if they find any relevant court decisions to their case (Dhanani, Mehta & Rana, 2021).

Such a tool can also support decision-making in legislation. For example, text-recognition tools can help monitor the implementation of European Union regulations in national legislation (Nanda et al., 2019). E-Discovery is an electronic tool for evidence searching in legal work, and text similarity analysis could also assist in this process by suggesting potentially relevant documents in a new case that are semantically similar to documents that were relevant in previous cases. Thus, by recognizing a semantic pattern, potentially relevant documents are immediately available, and no lengthy research work is required before the analysis process of e-Discovery (Trappey, Trappey & Liu, 2020).

In sum, semantic processing opens up new possibilities in legal research. Legal data can become more structured and expanded with more valuable metadata. It improves and automates the work of legal database editors, allowing them to do more value-added work. On the one hand, legal users
can speed up their searches. On the other hand, they can analyze individual legal documents and the relationships between documents much more comprehensively, allowing them to rely on more data to make their decisions.

However, in many cases, legal database developments using machine learning can be a significant burden on a company because it is not always known what kind of an AI solution is needed and how the providers could compare the different solutions. Thus, the last section presents the importance of text-processing pipelines.

4. Legal text-processing pipelines

There are a lot of AI- and NLP-based solutions that legal database providers can use to make their documents structured, valuable, and easily searchable. But it is also important that they need to customize processing their documents and their texts in their most optimal way, considering all their legal, linguistic, statistical, or other characteristics because different characteristics and regularity of the text need to be used depending on whether the task is categorization, summarization, or extraction.

However, legal documents are mainly stored in unstructured formats that are unsuitable for reading by algorithms and cannot be used to develop any NLP solution (Koniaris, Papastefanatos & Anagnostopoulos, 2018). This is what text-processing pipelines can help with. Unfortunately, the legal domain has a specific language and text structure, so general pipelines are not applicable in most cases (McCarty, 2009). This is why there is a need for legal domain-specific text-processing pipelines.

In short, a text processing pipeline is a tool that can convert unstructured text data into strings and can manipulate the text in the way computers, software, and algorithms can understand it. The main tasks that they usually do are sentence segmentation, word tokenization or vectorization, text lemmatization, and even others. We can say that these pipelines are the tools through which the abovementioned named entity recognition and semantic processing solutions can integrate into a concrete NLP project.

In the last 10 or 15 years, some NLP pipelines and frameworks have been proposed in the field of legal technology, for example, GarNLP (Bordino et al., 2021) and LexNLP (Bommarito II, Katz & Detterman, 2021). However, most of these solutions are language-dependent and can be used only in some
specific subdomain and for specific tasks. Another example of these pipelines is digital-twin-distiller. It is an open-source framework that can support the entire workflow of various NLP developments, and it can integrate multiple machine-learning algorithms, make different methods more comparable, and help reinforce each other's strengths (Csányi et al., 2022). By developing a legal database, providers can combine several parallel processing pipelines and replace the different AI methods as desired at any time. This way, the processing of legal documents is widely customizable, supporting their processing in their most optimal way, considering all their characteristics.

In an NLP development, the results of the different options for solutions must be comparable. For example, in a document categorization task, many pre-processing and vectorization methods can be used for different texts. It is also a point of decision whether we use any named entity recognition or semantic processing solution, and if so, which. Because, for example, the extraction of legal references and administrative bodies and the normalization of their names can help the efficiency of different algorithms. To select the best solution for the given task in terms of a given data set, these NLP methods and their results need to be compared. The digital-twin-distiller can help this process by allowing easy replacement of different components, and thus the results can be compared more quickly and efficiently. Thus, artificial intelligence-based support for legal and other expert activities can become more effective.

5. Conclusion

Continental legal databases nowadays often struggle with information crises and a large amount of unstructured legal text, which makes legal research a time-consuming and labour-intensive task. Named entity recognition solutions can be used to increase the efficiency of legal research. The most obvious way to use these solutions is to retrieve legal information entities automatically. The other way to use NER for legal texts is by creating citation networks for legal text. Creating a citation network between legal documents opens up new possibilities in legal document analysis, and can improve the result list of the database. A major limitation of these solutions is that they can only be used for explicit content in texts.

1 https://github.com/montana-knowledge-management/digital-twin-distiller
Semantic processing solutions, on the other hand, do not only extract and make the information elements contained in the text searchable, but allow a deep analysis of the legal text as a whole. It helps to represent the meaning of the legal text better, even through context. The applicability of this type of solution in the legal field is primarily the automatic categorization of different documents, automatic summary generation and text similarity analysis. In sum, NER and semantic processing open new possibilities in legal research and can increase the value of legal data.

Text processing or NLP pipelines and frameworks can help integrate the positives of the different NLP solutions. Digital-twin-distiller is a framework that can integrate multiple machine learning algorithms and helps make the different methods more comparable and mutually reinforce their strengths, even named entity recognition or semantic processing solutions. Also, it can help develop legal databases in the era of digitized documents.

Renátó Vági is a PhD student in law at Eötvös Loránd University, Hungary. His main research areas are data-driven technologies and their impact on law and society. He is working as a legal engineer at MONTANA Knowledge Management Ltd., where he deals with legaltech and other text-processing projects.

References


