Natural Language Engineering **25** (1): 211–217.

doi:10.1017/S1351324918000475

Industry Watch
Law and Word Order: NLP in Legal Tech

ROBERT DALE
Language Technology Group
e-mail: rdale@language-technology.com

(*Received 4th November 2018*)

Abstract

The law has language at its heart, so it's not surprising that software that operates on natural language has played a role in some areas of the legal profession for a long time. But the last few years have seen an increased interest in applying modern techniques to a wider range of problems, so I look here at how natural language processing is being used in the legal sector today.

The Legal NLP Landscape

The application of natural language processing, and artificial intelligence more generally, in the legal profession is not a new thing. The earliest systems for searching online legal content appeared in the 1960s and 1970s, and legal expert systems were a hot topic of discussion in the 1970s and 1980s.[1] But the last few years have seen a significant upsurge of interest in the area, including, as you might expect, an increasing number of start-ups claiming to apply deep learning techniques in the context of specific legal applications.

For a recent project, I had to review how NLP was being used in what has become known as Legal Tech. It turns out to be a densely populated space: a website at Stanford lists 1084 companies 'changing the way legal is done'.[2] In reviewing such a landscape, it helps to have a map. Conveniently, the practice of the law is a well-structured activity, with point solutions being available for a number of specific tasks that a typical law firm faces. My take is that there are five areas of legal activity where NLP is playing an increasing role:

- Legal research: finding information relevant to a legal decision
- Electronic discovery: determining the relevance of documents to an information request
- Contract review: checking that a contract is complete and avoids risk

[1] See Richard Susskind, *Expert Systems in Law*, Oxford University Press, 1987.
[2] https://techindex.law.stanford.edu

- Document automation: generating routine legal documents
- Legal advice: using question-and-answer dialogs to provide tailored advice

These brief definitions hide some complexity, which I'll unwrap below as I consider each area in turn.

Legal research

Legal research is the process of finding information that is needed to support legal decision-making. In practice, this generally means searching through both statute (as created by the legislature) and case law (as developed by the courts) to find what is relevant for some specific matter at hand.

That's a key purpose for the neatly organised bookshelves of thick bound volumes you see obligatorily lining the walls of lawyers' offices in court-room dramas and talking-head interviews. However, said volumes are often referred to as 'dusty tomes' for a reason: poring over the pages on a law library desk has long been displaced by electronic search and retrieval mechanisms.

LexisNexis (then called simply LEXIS) first appeared in the early 1970s, initially offering full text search of Ohio and New York case law; and it just grew from there. By the late 1970s, lawyers were able to access the database using dial-up services from dedicated terminals via 1200 baud modems. By the late 1990s, the data was on the web. Today Lexis Nexis claims to have over 30TB of content. Westlaw, another big player in the legal database world, was also founded in the mid-1970s, and was acquired by Thomson Corporation (now Thomson Reuters) in 1996. Add Wolters Kluwer and Bloomberg Law and you have the four major established providers in this space. Most law firms will have subscriptions to some or all of these services.

Despite the fact that the major players are so well-established, however, a number of newer players have captured some market share by offering smarter technologies that aim to improve the precision and recall of searches, beyond what can be achieved using 'traditional techniques', which here amount to good old-fashioned Boolean search and hand-constructed indexes.

Clearly the quality of the results of searching depends significantly on posing the right queries. Both CaseText[3] and CaseMine[4] provide interfaces that let you find related material by uploading a passage or even an entire brief that provides context for the search, thus supporting 'query by document'. In each case, this functionality is augmented by a range of neat UI features that facilitate the search task. As well as sidestepping the need to labour over appropriately-detailed search queries, this also increases the likelihood that additional relevant material not found by typical queries will be located.

[3] `https://casetext.com`; founded 2013, funding US$20.8M. All funding data provided here is from Crunchbase.

[4] `https://www.casemine.com`; founded 2013, funding unknown.

Taking a slightly different approach, Ross Intelligence[5] (which uses IBM Watson) and vLex[6] (with a product called Vincent) offer natural language query interfaces, so that 'you can pose your research questions like you're talking to another lawyer'.

Of course, the big four have been quick to build out their own 'AI-powered' solutions. In July 2018, LexisNexis launched Lexis Analytics, a legal research tool which incorporates the acquisition of machine learning and NLP start-up Ravel Law, amongst others. More or less at the same time, Thomson Reuters launched WestSearch Plus, a new search engine that claims to use state-of-the-art AI.

Electronic discovery

Electronic discovery, or e-discovery, is the process of identifying and collecting electronically-stored information in response to a request for production in a law suit or investigation. Faced with the hundreds of thousands of files that might reside on a typical hard drive, a key issue here is separating that content into what's relevant (or 'responsive', in the terminology of the domain) and what's not. In a case around a recent patent dispute with Apple, Samsung collected and processed about 3.6TB, or 11,108,653 documents; the cost of processing that evidence over a 20-month period was said to be more than US$13 million dollars.[7]

Today, the battle for market share is around optimised techniques for categorising whether documents are relevant as quickly and efficiently as possible. This process is called 'technology-assisted review' ('TAR'), and was for a number of years a focus of activity in the TREC Legal Track.[8] As with legal research, traditional approaches involved keyword or Boolean search, followed by manual review. More modern approaches use machine learning for document classification, referred to as 'predictive coding' in the legal profession. You want to maximise both precision and recall, while keeping the effort involved (in terms of the number of documents a human has to annotate or review) to a reasonable level. There is some debate in the legal community about the pros and cons of various techniques, and in particular around what counts as a reasonable seed set, and whether passive or active learning is better, where the former involves random selection of documents for human tagging, and the latter involves deliberate machine selection of either uncertain or, alternatively, assumed-relevant examples.[9]

Probably the biggest player in this space is Exterro.[10] Their newest technology, called Smart Labelling, avoids the need for users to provide initial seed sets of human-tagged documents, selecting for review the most relevant documents from

[5] https://rossintelligence.com; founded 2014, funding US$13.1M.
[6] https://vlex.com; founded 1998, funding 4M euros.
[7] https://blog.logikcull.com/find-out-how-much-samsung-paid-for-ediscovery-in-its-case-against-apple
[8] https://trec-legal.umiacs.umd.edu/
[9] See http://www.nonrelevant.net/2014/07/random-vs-active-selection-of-training-examples-in-e-discovery/ and http://www.wlrk.com/webdocs/wlrknew/AttorneyPubs/WLRK.23339.14.pdf.
[10] https://www.exterro.com; founded 2004, funding US$100M. Exterro's blog is a useful source of information on e-discovery.

the outset of the review process. DISCO[11] has a similar deep-learning-based solution in its 'Prioritized Review' process.

Everlaw,[12] on the other hand, seems still be using an approach where an initial seed set (they suggest 200 documents) must first be tagged. UI features can be important differentiators: Relativity,[13] previously known as kCura, also provides a phone app so you can 'code documents on your commute or on the couch'. Lawyers are 24/7 too.

Interestingly, generic NLP providers are also moving into the area. OpenText has introduced an e-discovery platform called Axcelerate;[14] and SDL, known for its translation products and services, provides a Multilingual eDiscovery Solution, enabling access to foreign language case-related content via translation.[15]

Contract review

A common activity for lawyers is to review contracts, make comments and changes, and advise their clients on whether to sign or negotiate for better terms. The contracts in question can be relatively simple, such as non-disclosure agreements (NDAs), or very large and complex, stretching to many hundreds of pages.

Automated contract review systems can be used to review documents which are relatively standardised and predictable in terms of the kinds of content they contain. The process involves decomposing the contract into its individual provisions or clauses, and then assessing each of these, either to extract key information or to compare against some standard (which might just be the set of other instances of such contracts held by a firm). So, for example, a contract review system might indicate the absence of a clause covering bribery, or indicate that a clause covering price increases fails to specify a percentage limit.

Contract review may be at the level of the individual contract, or—say, in the case of due diligence for a corporate acquisition—it may involve reviewing thousands of contracts on file. In the latter case, the technology begins to also incorporate aspects of what has become known as legal analytics, aggregating information across the data set to detect anomalies and outliers, and producing charts or tables that make it easy to compare across documents.

Contract review has generated a significant amount of interest in the last few years. Early approaches once more used the presence of key terms and headings to guide information extraction, and it's likely that many offerings still make use of some proportion of rule-based technology; however, not surprisingly, pretty much all the recent entrants into the space are using more sophisticated machine learning techniques.

[11] https://www.csdisco.com; founded 2012, funding US$50.6M.
[12] https://www.everlaw.com/; founded 2010; funding US$34.6M
[13] https://www.relativity.com; founded 2001, funding US$125M.
[14] https://www.opentext.com.au/what-we-do/products/discovery/axcelerate
[15] https://www.sdl.com/software-and-services/integrations/solutions-for-ediscovery.html

Three of the largest players here are Kira Systems,[16] Seal Software[17] and LawGeex.[18] Kira provides pre-built models for around 500 common provisions covering a range of contract types; you indicate which are relevant for the contract being reviewed, and you can also build custom models for provisions not already catered for. Seal offers similar capabilities but adds a logic engine that lets you apply business logic to the data extracted from the contracts reviewed; LawGeex emphasises the ability to compare contracts against pre-defined company policies.

A typical strategy for newer and smaller entrants seems to be that of beginning by focussing on quite specific document types, such as NDAs, real-estate leases, and privacy policies, and then increasing the range of documents dealt with as the company gains customers and traction. Leverton,[19] which was spun out of DFKI, focusses primarily on real estate documents. Targetting companies with large real estate portfolios, it processes contracts in 20 languages. Other smaller players worth a look are eBrevia,[20] Eigen Technologies,[21] LegalSifter,[22] and Luminance,[23] but there are many many others.

Not surprisingly, generic text analytics companies are also attracted to this use case: see ABBYY Text Analytics for Contracts,[24] Ayfie Contract Analysis,[25] and OpenText Perceptiv.[26]

Both ContractProbe[27] and PrivacyPolicyCheck[28] have online demos that let you upload documents for review. These are much simpler than the products discussed above, but they give a flavour of what contract review applications can do.

Document automation and legal advice

There's a fuzzy boundary between document automation systems and legal advice applications, so I'll consider the two categories together.

Legal advisors are interactive systems which, based on a set of questions posed by the system, produce advice tailored to the circumstances and requirements of the user. In many cases, the output is a legal document of some kind, so legal advice often amounts to document automation.

Document automation systems, on the other hand, typically use some kind of fill-in-the-blanks templating mechanism that enables the creation of a legal document tailored to specific criteria. In some cases, the data required to generate the document is obtained via an iterative question-and-answer dialog: a chatbot, if you like. In

[16] https://kirasystems.com/; founded 2015, funding CA$65M.
[17] https://www.seal-software.com; founded 2010, funding US$43M.
[18] https://www.lawgeex.com; founded 2014, funding US$21.5M.
[19] https://www.leverton.ai; funded 2012; funding 15M euros.
[20] https://ebrevia.com; founded 2012, funding US$4.3M.
[21] https://www.eigentech.com; founded 2014, funding UKP13M.
[22] https://www.legalsifter.com; founded 2013, funding US$6.2M.
[23] https://www.luminance.com; founded 2003, funding US$13M.
[24] https://www.abbyy.com/en-au/solutions/text-analytics-for-contracts
[25] https://www.ayfie.com/products/extensions/contract-analysis/
[26] https://www.opentext.com.au/what-we-do/products/discovery/perceptiv
[27] https://www.contractprobe.com
[28] https://privacypolicycheck.ai

such circumstances, the document automation sytem has the same kind of interface as that provided by a legal advice system.

The most publicly visible legal advisor is DoNotPay, an interactive tool whose initial focus was to help members of the general public to appeal parking tickets.[29] The scope of the application has grown immensely since then; at the time of writing, the DoNotPay app supports 14 different uses cases, including fighting unfair bank, credit card and overdraft fees, getting refunds from Uber and Lyft when a driver takes a wrong turn, and claiming refunds for late package deliveries.[30]

DoNotPay was created by Joshua Browder, a student at Stanford, in response to his own parking ticket experiences; but law firms are also interested in offering legal advice systems. Automation has clear advantages here, making available legal services to those who might not otherwise be able to afford them or be willing to pay for them.

So, for example, Norton Rose Fulbright, an Australian law firm, released at the end of 2017 a chatbot for privacy law concerns.[31] Built using IBM Watson, the tool answers standard questions about data breaches. The firm has since extended the application to handle GDPR queries.

Neota Logic,[32] which has been around since 2010, provides a platform for creating expert advisors; the technology application is in fact much broader than this suggests, since it can also be used for workflow automation and related tasks, including document automation, which I turn to next.

Legal document automation applications have been around for a long time, and are arguably amongst the earliest commercial natural language generation systems. Some researchers would be loath to call these template-based systems NLG, but the reality is that the technology used is similar to that offered by the leading commercial NLG vendors today.

These systems typically work by gathering relevant data from the user, either via form-filling or via a question-and-answer session, as noted above. The accumulated data is then used in a rule-based manner to craft a tailored document, via a combination of conditional document assembly and template slot-filling.

The most well-known offering in this category is Thomson Reuters' Contract Express,[33] whose target market is law firms that want to increase efficiency. There are other prominent players, including Rocket Lawyer,[34] which is more consumer-focussed, and Neota Logic (mentioned above), who provide both a generic Intelligent Document Automation facility as well as a more end-user-oriented specific application call PerfectNDA.[35]

[29] https://www.theguardian.com/technology/2016/jun/28/chatbot-ai-lawyer-donotpay-parking-tickets-london-new-york
[30] https://www.artificiallawyer.com/2018/10/12/the-genius-of-donotpay-giving-you-what-is-already-yours/
[31] http://www.nortonrosefulbright.com/news/159704/norton-rose-fulbright-launches-first-australian-law-firm-chatbot-to-help-manage-data-breach
[32] https://www.neotalogic.com/
[33] https://legal.thomsonreuters.com.au/products/contract-express/
[34] https://www.rocketlawyer.com/; founded 2008, funding US$46.2M.
[35] https://www.neotalogic.com/solution/perfectnda/

More generally, a number of organisations position their document automation offerings in the access-to-justice space, making tailored legal documentation easily available to the general public. Two such examples are A2J Author[36] and HelpSelfLegal.[37]

There are also companies that offer products which aim to help with patent drafting: see Specifio[38] and TurboPatent.[39]

Final judgement

As in many other areas, the nature of work in the legal profession is under threat from NLP and AI more generally. In early 2016, Deloitte found that 39% of jobs in the legal sector stood to be automated in the following ten years.[40] Recently, McKinsey estimated that 22% of a lawyer's job and 35% of a law clerk's job could be automated.[41] And as is common in other areas, you'll often see a positive spin put on this, with the usual claims that 'the technology frees up workers to do more interesting things'. But the pros and cons of technology uptake are a point of occasionally heated debate in the profession, which Richard Tromans characterises as consisting of conservatives who want to hang on to the status quo and progressives who want change.[42] A major barrier to change, of course, is that the legal profession has traditionally operated on the basis of billable hours. In that context, if a technology increases efficiency, it also reduces what you can put on the clock. On the other side, disruption of traditional approaches is inevitable in the face of increasing demands from acess-to-justice movements across the world.

I'd say the jury's no longer out on this one.

If you'd like to receive a short and snappy weekly newsletter on what's happening in the commercial NLP world, sign up for *This Week in NLP* at `www.language-technology.com/blog`.

[36] `https://www.a2jauthor.org`
[37] `https://www.helpselflegal.com`
[38] `https://specif.io`
[39] `https://turbopatent.com`
[40] `https://www.legaltechnology.com/latest-news/deloitte-insight-100000-legal-roles-to-be-automated`
[41] `https://www.linkedin.com/pulse/how-much-what-lawyers-do-can-automated-look-new-research-peter-nussey/`
[42] `https://www.artificiallawyer.com/2018/10/16/the-politics-of-legal-tech-progressives-vs-conservatives/`