CodeX Machine Generated Legal Documents Project

Document automation (also known as document assembly) pertains to systems and workflows that assist in creating electronic documents. The text assembled by computers into new documents may be pre-existing (or "canned") or computer-generated on-the-fly. Document automation in legal is verging on mainstream, if not already in some practice areas. Stanford Law School's CodeX Techindex, for example, lists around 250 legal document automation companies.[1]

Document automation has many benefits in law practice in that it reduces labor needs for rote and mundane writing, it reduces time spent proofreading, and it reduces risks associated with human error. Many of the associated risks, such as confidentiality and data security, are similar as with most other technologies used in law practice.[2]

The types of content contained in legal documents can be categorized into three groups: bespoke writing, mechanical writing, or canned text.[3] Bespoke writing reflects the intellectual heavy-lifting performed by the attorney preparing the document. It often involves original analysis on unique facts and is driven by creativity, judgment, strategy, and experience. Bespoke writing is too nuanced and context dependent to be a good candidate for automation.[4] Instead, this is where attorneys will continue to provide their primary value-add to legal documents.[5]

The world's most advanced NLP/NLG system, by far, exists between your ears.[6] Expecting a machine to do bespoke legal writing is unrealistic with today's technology. For example, a recurrent neural network (RNN) can create never-before-seen text, but only in relatively short spans (under 100 words), which may look bespoke at first glance but might be weird and completely nonsensical some (or all) of the time.[7] That might be fine for entertainment,[8] but not legal documents. There is also RNN plus neural bag-of-words (BoW), which has been used for predictive text (e.g., Gmail's Smart Compose), but it is limited to just a few words.[9] Predictive text at the current state-of-the-art may accelerate bespoke writing by speeding up typing, but it really does not take away any material *thinking* from the attorney authoring the document.

---

[1] See, https://techindex.law.stanford.edu/
[2] One risk worth mentioning specifically is failing to properly use or leverage technology, e.g., improper reliance on technology due to misunderstanding of its actual capabilities. This can result in inefficiencies associated with otherwise unnecessary human-performed document revisions and/or tech costs being passed to clients does not reflect full value that could/should be captured.
[3] [CITE]
[4] [CITE]
[5] [CITE]
[6] [CITE]
[7] [CITE]
[8] See, e.g., https://www.thisworddoesnotexist.com/; https://www.atrixnet.com/bs-generator.html.
[9] [CITE]

Mechanical writing describes the rote and mundane parts of traditional legal writing projects. It is driven by convention and/or by satisfying document requirements. Mechanical writing must be accurate and complete, but does not require significant mental work. Examples might include propagating certain project-specific language throughout a document, completing blanks in canned text, listing well-known examples, describing well-known facts, defining terms, etc.

Many technologies exist today that can auto-generate mechanical writing for various legal documents.[10] These technologies mostly use advanced templatization and language manipulation, which can be learning-based, rules-based, or a combination.[11] The manipulation often involves performing linguistic operations on some pre-existing body of text to generate new text. Common linguistic operations include text transduction (e.g., propagating certain language throughout the document; copy and paste and massage), text extraction (e.g., locating and describing facts to provide context and support), and text generation (e.g., summarization, data-to-text).[12]

Canned text is predetermined language, e.g., boilerplate. While there is no new writing happening, there is still a labor cost associated with canned text in manually-prepared documents. For example, identifying and locating appropriate canned text (template repository, old related documents, complete or partial document, etc.), organizing canned text in the document, and adapting to the current project by completing variable or conditional text.[13]

Basic automation tools for canned text are not new. Mail merge for simple documents (i.e., static template text with fill-in-the-blanks) is nearly as old as the word processor.[14] A complex and more contemporary example, however, might be a contract assembly tool.[15] An attorney identifies a type of contract and the basic facts and terms. Standard-language clauses are then automatically assembled into a draft contract document.

The ratios of these three types of content for any given document will determine (1) whether the document is a good candidate for automation and, if so, (2) which document automation techniques are likely to be most effective.
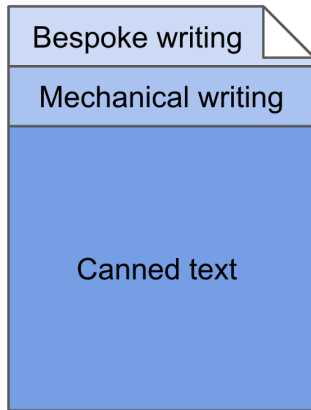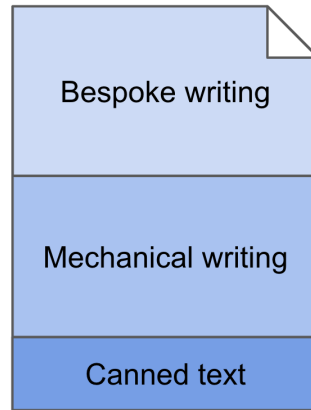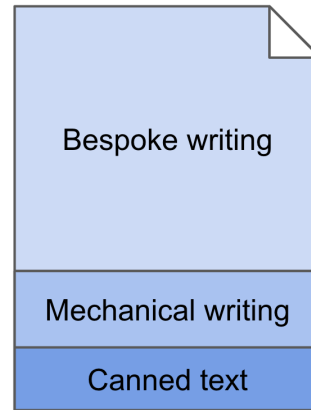
---

[10] [CITE]
[11] [CITE]
[12] [CITE]
[13] [CITE]
[14] [CITE]
[15] [CITE]

Contract	Brief	Memorandum
Bespoke writing	Bespoke writing	Bespoke writing
Mechanical writing	Mechanical writing	Mechanical writing
Canned text	Canned text	Canned text

Documents dominated by bespoke writing benefit the least from automation. From a technical standpoint, the best candidates for automation are ones with large portions being some combination of mechanical writing and canned text. Good technical candidates become good product candidates depending on the development costs to automate the document, document volume (e.g., docs/atty/year), market value of the document, and other factors.