Andy Tsao (adtsao), Vincent Sheu (vsheu), Jeff Sheng (jtsheng)
2017.10.26
CS 221: Artificial Intelligence: Principles and Techniques, Autumn 2017
TA Advisor: Steve Mussmann (mussmann)
Final Project Proposal

Define the input-output behavior of the system and the scope of the project.
On a broad level, we propose a system that takes documents as input, and outputs a modified version of those documents with functionally similar meaning, but with easier-to-comprehend syntax and/or vocabulary. We believe this has particular relevance in the legal world, where documents (e.g. contracts) are often horrendously long and contain arcane sentence structures and specialized terms of art.

What is your evaluation metric for success?
We plan on conducting surveys to evaluate our metric. Survey participants will be presented with a document in its unprocessed and processed state, and will be asked to rate each version's readability/comprehension (on a 0-9 or 1-10 scale, as well as a perceived educational level required to comprehend said document). As a control, we will run a separate cohort composed solely of law students, to ensure that a document's material terms have not changed.

Collect some preliminary data, and give concrete examples of inputs and outputs.
An example of a legal clause with a baseline translation, and a human (oracle) translation:

- Original text: "This Agreement contains the entire agreement between the Parties to this Agreement relating to the settlement and transactions contemplated hereby, and supersedes any and all prior agreements, understandings, representations, and statements between the Parties, whether oral or written, and whether by a Party or such Partys legal counsel. The Parties are entering into this Agreement based solely on the representations and warranties herein and not based on any promises, representations, and/or warranties not found herein. No modification, waiver, amendment, discharge, or change of this Agreement shall be valid unless the same is in writing."
- Baseline translation: This Accord contains the entire accord between the Celebrations to this Accord relating to the establishment and contracts contemplated hereby, and supersedes any and all prior accord, understandings, portrayals, and accounts between the Celebrations, whether oral or written, and whether by a Celebration or such Celebration's legal counsel. The Parties are entering into this Accord based solely on the portrayals and certificates herein and not based on any promises, portrayals, and/or certificates not found herein. No adjustment, postponement, change, discharge, or change of this Accord shall be valid unless the same is in writing.
- Oracle translation: "This Agreement contains the full, final agreement (including representations and warranties) between the Parties. It supersedes all previous discussions and agreements between the Parties, and can only be modified in writing."

Implement a baseline and an oracle and discuss the gap.
Baseline: Given a legal document, running a random replacement on all nouns sourcing from a thesaurus.

Oracle: Given a legal document, having a law student or lawyer “translating” the document into laymen’s plaintext.

The baseline is a fully-automated rudimentary process—assuming that words are only replaced with simpler words with the same meaning, the processed document should also have the same meaning (but with simpler vocabulary). The oracle performs what legal services can be modeled as today—having a trained professional (in our case, modeled by a fourth-year law student) doing an inefficient, manual deciphering. The gap in between rudimentary find-and-replace and intelligent deciphering represents a space that our project hopes to provide a result within—we don’t expect to be able to provide a perfect deconstruction of a legal document every time, but in this space, any amount of improvement would provide additional access to legal services.

What are the challenges?
Machine translation of text is already an imperfect science. Machine translation of legal documents (even within the same language) present additional problems. As a part of a field that requires three years of specialized graduate education and licensing exam(s) to join, legal documents and problems are particularly incomprehensible (see, e.g., every “Terms & Conditions” or “End User License Agreement” you’ve ever had to acknowledge to having read and understood, despite not actually having read and understood said document). Legal documents employ overly complex sentence structure and terms of art only familiar to those in the legal subspecialty (e.g. “person having ordinary skill in the art” in patent law, or negligently/recklessly/knowingly/purposefully in criminal law). In addition, an understanding of differing legal systems (e.g. legal precedent, rules of civil procedure, etc.) is often required to accurately interpret legal language.

Which topics (e.g., search, MDPs, etc.) might be able to address those challenges (at a high-level, since we haven't covered any techniques in detail at this point)?

- Machine Learning prediction
    - Supervised learning: Inputs of legal clauses, sentences, and documents and outputs of manual translations of said legal inputs as training data. This would consider n-grams of words via binary classification.
    - Unsupervised learning (NLP): Inputs of as many legal documents as possible, generating structure (e.g. word clusters of similar meaning, and conducting find-and-replace within word clusters to translate documents)
- Search: essentially machine translation, but the goal is to output text that's in the same language, but of simpler structure than the given input text. The output can be built out of actions, each action appending a word or a phrase to the current output.

Search the Internet for similar projects and mention the related work.
A search found many companies offer human translation services in the legal space, as well as at least one company (VIA) offering machine translation between languages (a space that Google Translate also operates in). However, we could find no results for other projects or companies attempting to use automated processes to make legal documents more readable.