Semantic Edge Labeling

- Citations, when a certain statute is being cited in another statute, differ in meaning, and we aim to annotate each edge with a semantic label that expresses this meaning or purpose. Our efforts involve defining, annotating, and automatically assigning each citation edge with a specific semantic label. We define a golden set of labels that cover a vast majority of citation types that appear in the United States Code (US Code) but still specific enough to meaningfully group each citation. We proposed a Linear-Chain CRF based model to extract the useful features needed to label each citation. The extracted features were then mapped to a vector space using a word embedding technique and we used clustering methods to group the citations to their corresponding labels. This paper analyzes the content and structure of the US Code, but most of the techniques used can be easily generalized to other legal documents. It is worth mentioning that during this process we also collected a human-labeled dataset of the US Code that can be very useful for future research.

Visual Grph Analysis

The UI was designed to allow someone to navigate the citation graph, examining each node and its subgraphs with a breadth-first search. It also has a function to search for a node, view the context in which the node was cited, and explore cycles that were pre-computed.

Cycles were selected as a particularly interesting case to test the UI on, as the number and types of cycles in the US Code was something that had not yet been reported on. 9,269 cycles were precomputed and integrated into the UI for a user to view. An example to the right shows a case that could be of interest to a user examining the robustness of the code, where every node is connected in a definition to another node.

The UI implementation was created with MongoDB, Express.js, and Cytoscape.js, a JavaScript graph visualization library.

Summery of Results

1. We proposed a label set long enough to cover almost all of the citations also short enough for practical use.
2. We trained and evaluated a Linear Chain CRF for predicate extraction (see Table).
3. Collected a dataset of 400 manually annotated citations.
4. Designed multiple automatic labeling schemes based on the extracted predicates.
   a) K-NN: 63.2%
   b) K-means: 61.6%
   c) Multiclass SVM (1-vs-1): 63.3%
   d) Human: At more than 71%

The individuals are trained and have special background in law.

Entity Extraction

NLP techniques applied to extract key information for each case. The entities extracted were the judges names, legal firm partners, locations, organization names and the orders passed by the judges.

Case Prediction

The length of the case were predicted with the KNN approach considering the first 25 weeks of the case to come up with an estimate for the completion of the case. The other model we employed is the KNN approach with the file patterns and the number of judges involved in the cases.

- K=6: Error: 52.6140
- MAPE: 54%
- K=6: Error: 52.6140
- MAPE: 46.4096
- MAPE: 48%

CONCLUSIONS

In the citation graph we have achieved setting up the golden labels and verified the thoroughness of the labels with the help manual annotators. We also apply many machine learning paradigms like SVM's, CNN to achieve the state of the art accuracy for the labelling tasks.

For the case prediction, we have extracted many features that are hard to extract from the documents. We successfully designed a model for the length of the case predictor.