

Progress in Language Processing Technology for Electronic Rulemaking

Stuart Shulman
University of Pittsburgh
121 University Place, Suite 600
Pittsburgh, PA 15260
shulman@pitt.edu

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213-8213
callan@cmu.edu

Eduard Hovy
USC-ISI
4676 Admiralty Way
Marina del Rey, CA 90292-6695
hovy@isi.edu

Stephen Zavestoski
University of San Francisco
2130 Fulton Street
San Francisco, CA 94117-1080
smzavestoski@usfca.edu

ABSTRACT

In this project, we are developing new text processing tools that help people perform advanced analysis of large collections of text commentary. This problem is increasingly faced by the U.S. federal government's regulation writers who formulate the rules and regulations that define the details of laws enacted by Congress. Our research focuses on text clustering, text searching, near-duplicate detection, opinion identification, stakeholder characterization, and extractive summarization, as well as the impact of such tools on the process of rulemaking itself. Versions of a Rule-Writer's Workbench are being built by researchers at ISI and CMU, made available for experimental use by our government partners at the DOT and EPA, and evaluated by researchers at the Library and Information Science and Sociology departments at the universities of Pittsburgh and San Francisco, respectively. This project started in October 2004 and is funded for 3 years under the NSF's Digital Government Program.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Content Analysis and Indexing – *indexing methods, linguistic processing.*

General Terms

Algorithms, Experimentation.

Keywords

Information retrieval, information extraction, duplicate detection, opinion recognition, regulatory rulemaking, federal government.

1. MOTIVATION AND BACKGROUND

Many people today—including news analysts, opinion pollsters, advertisers, and government regulation writers—need to interpret, structure, and rapidly master large quantities of commentary about their work. We focus on the task facing several thousand regulation writers who formulate the rules and regulations required for the implementation of laws.

In this procedure, they invite and then process in detail comments from the public on their proposed regulations. In some instances, they may receive several hundred thousand form letters by email or studies of a few hundred or thousand pages.

A decade ago form letters were not difficult to process. Except for the signature, address and possibly a brief hand-written comment, they were exact duplicates of the original form letter. Form letters could be identified and sorted easily, so that the original could be considered, and the duplicative copies counted, reported, and then largely ignored.

Since our last report, there continue to be electronic mass comment campaigns. For example, the USDOT has a new Corporate Average Fuel Economy (CAFE) rule forthcoming that received over 40,000 public comments. As part of its continuing support for this research, the DOT expeditiously provided our group with this data for future experiments and hosting on the eRulemaking Testbed at CMU.

We have also been in a discussion with personnel at the Union of Concerned Scientists about its use of web-based form letter campaigns. This conversation seeks to advance the activist-business-government-researcher dialogue generated as by-product of the past 5 years of NSF-funded studies of electronic rulemaking.

This particular multi-year study will expand and upgrade the scope and function of CMU's eRulemaking Testbed and systematically feedback information from users in and out of government via workshops, focus groups, Web surveys, reports, publications, and presentations.

Our research explores the use of information extraction and information retrieval to develop tools that assist rule-writers and analysts in managing large volumes of public comments. Information extraction techniques strip off email headers, salutations, signature lines, and advertising text. Text clustering algorithms identify exact duplicates, group together comments that are similar but not identical, and organize them hierarchically for browsing by rule-writers. Text-differencing algorithms identify where a person has edited a form letter so that a rule-writer's attention is drawn immediately to the unique part of an edited form letter.

Our project aims to develop tools for interpreting, structuring, and rapidly mastering large quantities of opinion-based text. This study builds on our previously developed eRulemaking testbed by systematically collecting information from users of the testbed's text analysis tools. The testbed's new text processing tools perform text clustering, text searching using information retrieval, near-duplicate detection, opinion identification, stakeholder characterization, and extractive summarization of large volumes of public commentary.

2. WORK TO DATE

The social science component of the project convened a focus group in September 2005. The subjects were citizen commenters from an earlier Union of Concerned Scientists mass comment campaign who were chosen at random from across the U.S. The objective was to understand more thoroughly the motivations of mass campaign commenters, their decisions in using and modifying form letters, and the expectations they have regarding the way in which agencies respond to their comments.

A new Qualitative Data Analysis Program (QDAP), initiated by Dr. Stuart Shulman at the University of Pittsburgh, continues large scale manual coding of thousands of public comments and press accounts of the rulemakings under study. These manual annotation techniques were further refined in order to improve inter-coder reliability and to increase their utility to natural language processing researchers. We continue to experiment with training and annotation innovations and hope to host a workshop on this specific task in the near future. The challenge has been to develop tailored, reliable coding schemes that can serve both the social and computer science research communities.

Our previous near-duplicate detection work was extended significantly this year, resulting in **DURIAN (DUPLICATE Removal In lARge collectionN)**, an algorithm using a traditional bag-of-words document representation, document attributes ("metadata"), and document content structure to identify form letters and their edited copies in public comment collections. In Yang et al. (under review), we report results against two samples drawn from a large set of public comments submitted to the EPA during the summer of 2005. We plan further evaluation using corpora provided by another agency. Early tests suggest that DURIAN is about as effective as human assessors at detecting duplicates. Further analysis will provide greater insight into the strengths, weaknesses, and generality of the algorithm.

The second substantive accomplishment has been the introduction of the eRuleClient tool, through which QDAP coders can more easily annotate both the subtopic codes and argument linkage structures of public comments. Through several iterations with

QDAP coders implementing the EruleClient tool and researchers and programmers at ISI making on-the-fly updates, coder satisfaction with the tool was greatly enhanced.

Two further accomplishments are the creation of technology that categorizes each sentence into one of nine topic categories (e.g., Economic, Environment, Technology, etc.), and the deployment of opinion detection software, allowing the user to identify all fragments that are positive or negative about the overall theme. This engine achieves an F-score of 71% (Kwon et al., under review). Work underway seeks to determine the argument structure (main topic, subtopics, and dependencies) of the texts.

The next step of the research is to combine all the above work in order to create a system that performs multi-aspect analysis of rulemaking comments and provides a useful review tool for rulemakers, and then to test its efficacy with government users.

3. PRESENTATIONS IN 2005

We have presented this work at the following venues:

American Bar Association's Administrative Law Conference
 American Political Science Association
 European University Institute
 International Conference on E-Gov Research
 Judiciary Committee Symposium, U.S. House of Representatives
 National Association of Secretaries of State
 National Conference on Digital Government
 Second Conference on Online Deliberation
 State Department Speaker/Specialist Tour of Kazakhstan

4. PROJECT HOME PAGE AND TESTBED

For more information:

<http://erulemaking.ucsur.pitt.edu>

The eRulemaking Testbed:

<http://hartford.lti.cs.cmu.edu/eRulemaking/Data.html>

5. ACKNOWLEDGMENTS

We thank personnel at the USDA, DOT, and EPA for providing the public comment data and insights that made this research possible. This material is based on work supported by National Science Foundation (NSF) grants IIS-0429293, 0429102, 0429360, and 0429243. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors' and do not necessarily reflect those of the NSF.

6. REFERENCES

- [1] H. Yang and J. Callan. (2005). Near-duplicate detection for eRulemaking. *Proceedings of the Fifth National Conference on Digital Government Research*. Atlanta, GA.
- [2] H. Yang, J. Callan, and S. Shulman. (Under review). Next steps in near-duplicate detection for eRulemaking. *Proceedings of the Sixth National Conference on Digital Government Research*. San Diego, CA.
- [3] N. Kwon, S. Shulman, and E. Hovy. (Under review). Collective text analysis for eRulemaking. *Proceedings of the Sixth National Conference on Digital Government Research*. San Diego, CA.