Language Processing Technologies for Electronic Rulemaking: A Project Highlight

Stuart Shulman
University of Pittsburgh
121 University Place, Suite 600
Pittsburgh, PA 15260
+1-412-624-3776
shulman@pitt.edu

Eduard Hovy
USC-ISI
4676 Admiralty Way
Marina del Rey, CA 90292-6695
+1-310-448-8731
hovy@isi.edu

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213-8213
+1-412-268-4525
callan@cmu.edu

Stephen Zavestoski
University of San Francisco
2130 Fulton St.
San Francisco, CA 94117-1080
+1-415-422-5485
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 United States 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 using information retrieval, 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 will be built by Computer Science researchers at ISI and CMU, deployed annually for experimental use by our government partners, and evaluated by social science researchers from the Library and Information Science and Sociology departments at the Universities of Pittsburgh and San Francisco respectively. This three-year project started in October 2004 and is funded under the National Science Foundation's Digital Government program.

Categories and Subject Descriptors

H.3.3 INFORMATION SEARCH AND RETRIEVAL

General Terms

Algorithms, human factors

Keywords

Natural language processing, information retrieval, near-duplicate detection, opinion recognition, text annotation, regulations, rulemaking, federal government

1. MOTIVATION AND BACKGROUND

Since its formation, the eRulemaking Research Group has participated in and organized workshops, issued a focus group-based stakeholder report, and made presentations to federal agencies as well as an international academic community. We launched an eRulemaking text data testbed, created a group web site that has had over 18,000 pages of written and presentation content viewed or downloaded since its late May 2004 inception, and we collaborated with five federal agencies in the successful submission of a proposal to the NSF's Digital Government program. This multi-year study expands and upgrades the scope and function of the eRulemaking testbed and systematically feeds back information from users in and out of government via workshops, focus groups, Web surveys, reports, publications, and regular national and international presentations.

Most proposed regulations attract relatively few comments, but high profile regulations can attract hundreds of thousands of comments from the public. When the volume of comments is large, it is usually because one or more stakeholder organizations publicized the proposed regulation and created a letter template (a "form letter") to which people can attach their name.

A decade ago form letters were not difficult to process. Except for the signature, sender's 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.

Now, agencies accept comments via e-mail and the Web. With the assistance of an expanding array of for-profit web services, it is easy (though not cheap) for interest groups to ask members of the public to modify a form letter to reflect their particular viewpoint. As a result, the task of identifying distinct and substantive viewpoints is now dramatically more difficult. Agencies have experimented with simple heuristics, for example based on the length of a comment or its email routing path, but manual sorting, often by private sector contractors, remains the predominant method. For example, the Environmental Protection Agency (EPA) recently printed over 500,000 emails on their proposed mercury regulation and had 15 staff members sort by sight the duplicates, triplicates, and all manner of variants, as well as the odd substantive and original non-form letter comment.

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 comments that are similar but not identical, and organize them hierarchically for browsing by rule-writers. Text-differencing algorithms identify where a form letter has been edited 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 will build 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 will perform text clustering, text searching using information retrieval, near-duplicate detection, opinion identification, stakeholder characterization, and extractive summarization of large volumes of public commentary. These tools will be deployed as prototypes of components that might be used in a Rule-Writer's Workbench, and will be evaluated by our government partners as well as other interested parties, such as stakeholder groups and web advocacy designers.

2. WORK TO DATE

The social science component of the project has focused on human coding and analysis of public commentary. During 2003-2004, a team of ten undergraduate students (funded by the NSF, SES-0322662) coded a total of 1,500 comments submitted during three different federal rulemaking proceedings. A new team of graduate student coders, also working under this funding, is currently building on the coding techniques developed last year by applying them to new 1000 comment random samples drawn from a dataset of over 500,000 emails submitted to the EPA about the new mercury air pollution regulation.

Using two rounds of coding the same documents, with two coders reviewing the work of each individual coder, and coder overlap on 32% of the documents, we have established an average .8 kappa coefficient score on our inter-rater reliability. We continue to experiment with training and annotation innovations. The challenge here is to develop tailored, reliable, robust coding schemes that can serve the analytical needs of both the social and computer science research communities.

Our preliminary computer science research has been tested on several public comment datasets supplied by the United States Department of Agriculture (USDA), the Department of Transportation (DOT), and the EPA. The EPA's email-only public comment dataset from the proposed "National Emission Standards for Hazardous Air Pollutants for Utility Air Toxics" rulemaking, colloquially known as the "Mercury" dataset, consists of 536,975 email messages. We provide Web-based search and browsing capabilities so that rule-writers can try out the algorithms and suggest changes.

As is common in most research projects, our initial graphical user interfaces were simple and designed for use by experts. Rule-writers identified this as an obstacle to serious evaluation, so we now invest more effort in understanding human-computer interaction issues. We have also developed "ground truth" labeling of some of the data, so that we can conduct rigorous quantitative evaluations [Yang and Callan, 2005].

The refined coding process and structure will assist with the development of automated text processing tools. Comparisons of human and automated coding will help show where computers make the routine tasks more manageable and accurate, and where the human eye is still indispensable to sorting and analysis.

The computer science component of the project is developing new text processing tools that can perform advanced analysis of such collections, including text clustering, text searching using information retrieval, near-duplicate detection, opinion identification, stakeholder characterization, and extractive summarization. A preliminary near-duplicate detection algorithm has been developed and tested at CMU, and forms the basis of further developments. Research at ISI is focusing on the automated detection of opinions within comments, and the construction of 'argument graphs' that track the commenter's reasoning and opinion-support argumentation.

In the future we will explore more sophisticated analysis and organization of the comments, for example by automatically identifying stakeholders, key issues, and opinions mentioned in messages.

3. PRESENTATIONS

We have presented this work at the following venues:
American Bar Association's Administrative Law Conference
American University Rulemaking Workshops
American Political Science Association
Council of Europe "The Future of Democracy in Europe"
Department of Transportation
Environmental Protection Agency (EPA)
EPA's Office of Environmental Information National Meeting
Institute of Public Administration of Canada
National Association of Secretaries of State
Oxford Internet Institute
United States Chamber of Commerce Regulatory Affairs
Western Political Science Association

ACKNOWLEDGEMENTS

We wish to thank personnel at the USDA, DOT, and EPA for providing the public comment data and the insights that make this research possible. This research was supported by NSF grants IIS-0429293, 0429102, 0429360 & 0429243. Any opinions, findings, conclusions, or recommendations expressed in this paper are the authors' and do not necessarily reflect those of the National Science Foundation.

REFERENCES

- H. Yang and J. Callan. "Near duplicate detection for eRulemaking." In *Proceedings of the Fifth National Conference on Digital Government Research*. 2005.
- The eRulemaking Research Group project home page is at: http://erulemaking.ucsur.pitt.edu
- The eRulemaking Testbed is at: http://hartford.lti.cs.cmu.edu/eRulemaking/Data.html