Using Natural Language Processing to Improve eRulemaking

[Project Highlight]

Claire Cardie
Information Science Program
and Department of Computer Science
Cornell University
Ithaca, NY USA
cardie@cs.cornell.edu

Cynthia Farina
Law School
Cornell University
Ithaca, NY USA
crf7cardie@cornell.edu

Thomas Bruce
Legal Information Institute
Cornell University
Ithaca, NY USA
trb2@cs.cornell.edu

ABSTRACT

This paper describes in brief Cornell's interdisciplinary eRulemaking project that was recently funded (December, 2005) by the National Science Foundation.

1. INTRODUCTION

Each year federal regulatory agencies issue more than 4,000 new rules [6]. By law, many of these must be created through a complex and expensive process in which the agency drafts a proposed rule and then exposes the proposal, any underlying data, and its legal and policy rationale to public comment. This process, *notice and comment (N&C) rulemaking*, is the mechanism through which most agencies make major regulatory policy. One of, if not the, most important functions of government agencies [6, 5][1], N&C rulemaking is also one of the slowest. A duration of two to five years is not uncommon [5][2].

In N&C rulemaking, the agency may receive anywhere from dozens, to hundreds of thousands, of comments, depending on the subject and complexity of the rule. The agency's fundamental legal obligation is to review all the comments received and, if it chooses to adopt the proposed rule, to issue a statement that not only (i) demonstrates why its choice is within its statutory authority and sound as a matter of regulatory policy, but also (ii) responds to significant criticisms made in the comments and explains why it rejected alternative approaches suggested there [10][3]. The stakes for the agency are high. Failure to adequately address critical comments and discuss alternatives in the statement accompanying the final rule can lead a court to invalidate the rule thereby requiring still more agency time and effort to perform additional review and explanation.[10][4]

The need to absorb and assess the significance of hundreds, or even thousands, of comments is not the only hurdle that confronts the agency trying to make regulatory policy through N&C rulemaking. Over the last 25 years, Congress and the President have imposed an increasing number of mandates on rulemaking regardless of regulatory subject area [9]. These mandates are typically designed to protect a specific interest (such as small businesses or Native American tribes) or are triggered when a proposed rule would pass a certain threshold (such as a certain dollar amount of economic impact). They may require that, before completing the rulemaking, the agency prepare a certain kind of analysis, consult with another agency or a particular private entity, or issue a specified certification. Rule writers have found it increasingly difficult to keep track of these mandates and to recognize which, if any, are relevant in a particular rulemaking [7, 9]. As a result, they may complete the long and expensive N&C process only to discover that an arcane but legally required assessment, consultation, or certification was triggered but not accomplished.

Electronic rulemaking (eRulemaking) includes a wide range of ways that information technology might be used in rulemaking. It includes, but is not limited to: converting the agency's docket (the filing system showing all its activities, including rulemaking) to electronic form and making it available via the Internet; allowing submission of comments via email and the Internet in addition to (and perhaps eventually in place of) conventional mail and fax; and using search engine, hypertext, and other IT capacities to allow both the public and agency rule writers to find, sort, and link the massive amount of material relevant in a rulemaking more easily and cheaply than could possibly be done with hard copies.

eRulemaking thus has the potential to radically transform the N&C process. It could make the process more transparent and accessible to the public, and more substantively

[1] At 180,280-83.
[2] At 102-04.
[3] At 524-50.
[4] At 524-50 & 1016-26.

reliable and cost-effective for the agency.

To be sure, Module III of the eRulemaking Initiative contemplates developing a "rule writer's tool kit" to help categorize comments, mine data, and provide online rulewriting instruction. While existing language processing techniques (e.g. for information retrieval, text categorization, document clustering, and information extraction) could provide some of the basic capabilities listed above, they would require significant testing and evaluation within the eRulemaking domain. In addition, research on methods that would clearly be invaluable in actually carrying out the more complex of these tasks has only barely begun. Work in the area of text summarization and sentiment analysis, for example, is still very new [8, 1, 2, 11], but will be essential to analyze and summarize the opinions expressed in comments.

2. PROJECT GOALS

Our propose to apply and develop a range of methods from the field of natural language processing (NLP) to create NLP tools to aid agency rule writers in:

- organization, analysis, and management of the sometimes overwhelming volume of comments, studies, and other supporting documents associated with a proposed rule; and
- analyzing proposed rules to flag possibly relevant mandates from the large number of statutes and Executive Orders that require studies, consultations, or certifications during rulemaking.

Officials from the Departments of Transportation and Commerce, with whom we are collaborating in the project, identified both tasks as high priority needs. We will focus on the use of information extraction, text categorization, and opinion-oriented text analysis techniques in both supervised and weakly supervised machine learning frameworks. Importantly, we will also focus on the use of human language technologies to elicit more informed comments from commenters. The tools and methods we develop should be valuable not only in the eRulemaking arena, but also in business (e.g. automatic analysis of online product reviews), government intelligence (e.g. analyzing emerging opinion on a hot topic in the Mideastern vs. European press), science (e.g. extracting information from biomedical literature to create a database), and social science (e.g. processing Weblogs).

We will evaluate the integration of the tools into the day-to-day rulemaking process by applying qualitative and quantitative methods from social sciences — survey instruments, longitudinal interviews, and direct observation [4].

More generally, we will study the effect of technology on the rulemaking process. Despite the crucial importance of rulemaking to federal regulatory policymaking, there is a serious shortage of research on how the process actually occurs within agencies [5, 3][5].

3. PLANS FOR 2006

Our plans for 2006 include a number of related efforts, each of which aims to proactively use technology, usually human language technology, to improve eRulemaking for rule writers and for the public:

- Begin the creation of an eRulemaker's "best practices" guide.
- Investigate options for providing technical support for the creation of hyperlinks between (parts of) a proposed rule and relevant law.
- Develop ways to streamline the process of educating the public on the process and substance of rulemaking.
- Investigate options for employing NLP techniques to elicit better comments.

[5]Kerwin at 279-83.

4. ACKNOWLEDGMENTS

This work is supported in part by NSF Grant IIS-0535099 and by a Xerox Foundation and a Google gift to the first author.

5. ADDITIONAL AUTHORS

Additional authors: Erica Wagner (School of Hotel Administration, Cornell University) email: `elw32@cornell.edu`.

6. REFERENCES

[1] C. Cardie, J. Wiebe, T. Wilson, and D. Litman. Low-level annotations and summary representations of opinions for multi-perspective question answering. In M. Maybury, editor, *New Directions in Question Answering*. 2004.

[2] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In *HLT-EMNLP 2005*, 2005.

[3] C. Coglianese. The state of rulemaking in the federal government. Technical report, Transcript Panel 6, 2005.

[4] B. Kaplan and D. Duchon. Combining qualitative and quantitative methods in information systems research: A case study. *MIS Quarterly*, 12(4), 1988.

[5] C. Kerwin. The state of rulemaking conference. Technical report, Transcript Panel 1 and 6, 2003.

[6] C. Kerwin. The state of rulemaking in the federal government. Technical report, Transcript Panel 1, 2005.

[7] T. O. McGarity. *The Expanded Debate Over the Future of the Regulatory State. 63 U. Chi. L. Rev. 1463, 1523.* 1996.

[8] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP-2002*, pages 79–86, 2002.

[9] M. Seidenfeld. *A Table of Requirements for Federal Administrative Rulemaking. 27 Fla. S.U.L. Rev. 533, 535.* 2000.

[10] P. Strauss, T. Rakoff, and C. Farina. *Administrative Law.* 10th edition, 2003.

[11] L. Zhou and E. Hovy. Digesting virtual "geek" culture: The summarization of technical internet relay chats. In *ACL-2005*, 2005.