Query-based Opinion Summarization for Legal Blog Entries

Jack G. Conrad, Jochen L. Leidner, Frank Schilder, Ravi Kondadadi
Research & Development
Thomson Reuters Corporation
St. Paul, MN 55123 USA
{Jack.G.Conrad, Jochen.Leidner, Frank.Schilder, Ravikumar.Kondadadi}@ThomsonReuters.com

ABSTRACT

We present the first report of automatic sentiment summarization in the legal domain. This work is based on processing a set of legal questions with a system consisting of a semi-automatic Web blog search module and *FastSum*, a fully automatic extractive multi-document sentiment summarization system. We provide quantitative evaluation results of the summaries using legal expert reviewers. We report baseline evaluation results for query-based sentiment summarization for legal blogs: on a five-point scale, average responsiveness and linguistic quality are slightly higher than 2 (with human inter-rater agreement at $\kappa = 0.75$). To the best of our knowledge, this is the first evaluation of sentiment summarization in the legal blogosphere.

Categories and Subject Descriptors

H.3.0.a [**Information Storage and Retrieval**]: General—*Question Answering and Summarization*; I.2.7.i [**Information Storage and Retrieval**]: Natural Language Processing—*Opinion Mining*

General Terms

Modeling, Experimentation, Measurement

Keywords

opinion detection, sentiment analysis, summarization, question answering, query log analysis, legal informatics

1. INTRODUCTION

In today's dynamic legal information environment, legal professionals are increasingly compelled to extend the boundaries of their quest for sources of relevant information and knowledge of legal trends. For this reason, a growing number of professionals are taking an interest in legal blogs (a.k.a. "*blawgs*") that increasingly provide useful information, information that is typically composed by legal scholars, lawyers or students of law. The topics discussed in blogs are of special interest, since they often express the affective state (opinion) of the blog owner with respect to a subject. Compiling blawg entries with respect to a particular legal topic and summarizing them for these professionals could potentially be of great value to them. Not only could such a resource potentially support legal research, but it could also assist in making business decisions tied to the outcomes of pending legal decisions being discussed by legal professions in the blogosphere. A number of other compelling applications of such opinion mining or "sentiment analysis" of legal blogs are described below.

ICAIL '09, June 8-12, 2009, Barcelona, Spain

Prospective Applications. There exist a number of practical applications regarding opinion mining of legal blogs [6]. Some of these include:

- *profiling* — representing litigation patterns of various attorneys or issue-specific dispositions of certain courts;
- *alerting* — informing legal subscribers of unfavorable news and disclosures that may impact the clients of a firm;
- *monitoring* — following what communities are saying about certain firms, legal research products or services;
- *tracking* — studying decisions of judges or reputations of law firms based on client feedback over time;
- *hosting & surveying* — allocating blog space for practitioners to comment on legal topics and decisions that can subsequently be mined for trends;
- *exploration & education* — at law schools, harnessing legal sentiment summarization as a means of engaging students with contrasting legal opinions.

The degree to which these will become essential to the field of law will depend on technological, economic, and domain-related factors, not to mention parallel developments in other professional fields.

Opinion summarization task definition. The starting point of our research is the TAC opinion summarization task [9] which we modified for summarizing legal blogs. The 2008 Opinion Summarization task is defined as the automatic generation of well-organized, fluent summaries of opinions about specified targets, as found in a set of blog documents. Each summary has to address a set of complex questions about the target, where the question cannot be answered simply with a named entity (or even a list of named entities). The input to the summarization task comprises a target, some opinion-related questions about the target (see Figure 1), and a set of documents that contain answers to the questions. The output is a summary for each target that summarizes the answers to the questions.

Target: Windows Vista

Questions:

- What features do people like about Vista?
- What features do people dislike about Vista?

Figure 1: A TAC *topic* is a target along with associated questions.

Sidebar: Annual Evaluation Efforts in Text Analytics

In recent years the community of information retrieval and natural language processing researchers have been engaging their peers regularly to evaluate their methods and assess the progress of their research. We outline some of their activities here, and invite and encourage the AI & Law community to consider joining these efforts.

TREC. The Text REtrieval Conference ("TREC") is an annual international workshop carried out by NIST, the U.S. National Institute of Standards and Technology [27], where the organizers disseminate information retrieval tasks and datasets, and participants develop systems that solve the tasks and submit their results to NIST for evaluation. Participants can also propose new tasks for subsequent years. TREC has had a legal track since 2006 (see http://trec-legal.umiacs.umd.edu/ for details).

DUC. The Document Understanding Conference (DUC) [19] was a workshop running from 2001-2007, in which a large number of automatic summarization systems have been evaluated (see http://duc.nist.gov for details).

TAC. In 2008, NIST created the first Text Analysis Conference (TAC) in order to give the natural language processing community a platform serving similar purposes that TREC serves for IR, and to consolidate different sub-communities (TREC, DUC, and RTE, the PASCAL Textual Entailment Challenge) (for further details, see http://www.nist.gov/tac/).

Legal text analytics systems can be evaluated using the same paradigms used at TAC/TREC, and using the same dataset allows for better comparisons of results.

In this paper, we describe the first attempt to apply an open-domain document sentiment summarization system to the legal blogosphere. We define a set of legal questions asking about a sentiment-related topic from the legal realm, and retrieve the top-10 documents from BlogSearchEngine,[1] a leading blog search engine not limited to the legal profession, but focusing on high quality blogs. We process the resulting set of documents with **FastSum** [23], a summarization system recently enhanced to produce sentiment summaries [24, 25]. We describe our research efforts, discuss some limitations we encountered, and present preliminary evaluation results based on human ratings of the summaries performed by legal professionals.

In particular, we modify the evaluation guidelines of the first Text Analysis Conference (TAC) outlined above to better evaluate the accuracy of the summaries with respect to the polarity of the question (see sidebar above). We also describe a proposal to the AI & Law community to consider a *blawg legal summary track*, for instance, at a forthcoming TAC conference.

Paper Structure. The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 provides a system description of our summarization application. Sections 4 and 5 describe our data and methodology, respectively. In Section 6, we report on the results of our evaluation and their significance. Section 7 proposes a research community-sponsored shared task. Section 8 lays out our conclusions, while future work is addressed in Section 9.

2. RELATED WORK

Lerman *et al.* performed an evaluation that shows that users have a strong preference for summarizers that model sentiment over non-sentiment baselines [14]. Ho and Quinn measure the political orientation of 25 U.S. newspapers, using 1,500 editorials from 1994-2004 [12]. Unlike the present approach, their method rates the newspapers in general, not with respect to a particular input query. Ashley and Aleven present an intelligent tutoring system for teaching law students to argue with cases [1]. Their objective is educational, while we aim to build a system that is robust enough to process evidence on legal sentiment from the Web, in order to support the professional researcher. Hachey and Grover apply *argumentative zoning* to the task of summarizing legal decisions by the House of Lords [11]. Unlike their work, we do not use a zoning technique, and our summaries are query-based. LetSum, a similar system to Hachey and Grover's system, was created by Farzindar and Lapalme [10]. This system summarizes Canadian court decisions. It is remarkable in that it identifies roles of text spans. Saravanan, *et al.* [22] report on legal summarization using conditional random fields (CRFs), i.e., undirected statistical graphical models. They report a high F-score (> 0.8) and also consider the extraction of rhetorical roles (e.g. title, petitioner, respondent) of text spans. In contrast with the above legal summarization work, our system works with highly heterogeneous Web-based blog entries in response to a user's information need that is generally expressed along a particular sentimental dimension (e.g., positive view towards subject vs. negative view towards subject).

Quaresma and Rodrigues present a system for legal Q&A in Portuguese [20]. A parser is used to create a representation according to the Discourse Representation Theory (DRT) of the input in a PROLOG-like language. But the authors do not address sentiment, nor are performance figures of their Q&A system reported on Web blog data.

The U.S. National Institute of Standards and Technology (NIST) has been conducting regular annual open evaluations in legal retrieval and blog retrieval at the Text RE-

[1] www.blogsearchengine.com

trieval Conference (TREC) [27]. However, the "blog track" did not address legal sentiment summarization, but addressed ad-hoc document retrieval, while the "legal track" addressed retrieval for e-Discovery rather than summarization [18].

Conrad and Schilder [6] examine a small legal blog test collection and present first results on language model based classification with respect to subjectivity and sentiment polarity. In the context of summarization, Schilder *et al.* [24, 25] present a first query-based sentiment summarization system that produces a gist of the sentiment found in blogs and news. This work is similar. The earlier work uses questions of general interest, whereas the current work focuses on legal questions.

To the best of the authors' knowledge, this paper presents the first approach to summarize legal sentiment with respect to a legal question based on Web log ("blawg") evidence.

3. SYSTEM

FastSum [23] is a multi-document summarization system that was subsequently modified for sentiment. It uses a regression SVM for training a sentence classifier for good summary sentences similar to [15]. An important part of FastSum is a filtering component that identifies sentences that are unlikely to be in a good summary (e.g., no word overlap between query and sentence, difference in length). Another filter is concerned with the sentiment of a sentence. We added this filter in order to carry out opinion summarization.

The overall workflow of the FastSum blog opinion summarization system, as described in [24], is illustrated in Figure 2. At a high level of abstraction, the principal components of the FastSum blog opinion summarizer are shown in sequential order:

A. Pre-processing;

B. Question sentiment and target analyzer;

C. Filtering;

D. Feature extraction;

E. Sentence ranker;

F. Redundancy removal;

There are particular components we would like to highlight in this system overview (portions shared in Figure 2), because we made crucial changes to these components in order to allow for the summarization of legal blogs:

A.1 HTML parsing and clean-up module;

B.1 Question sentiment and target analyzer;

C.1 Sentiment tagger;

C.2 Target overlap;

We briefly discuss why finding the target for a legal question is more difficult than identifying the target for other sentiment questions, as used in the TAC data.

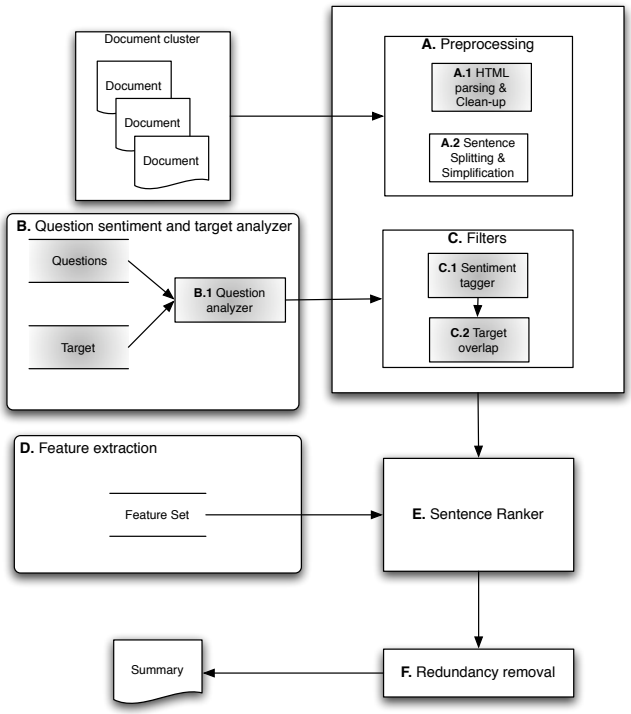

Figure 2: FastSum architecture for blog opinion summarization.

3.1 Pre-processing, query analysis and filtering components

The pre-processing module carries out tokenization and sentence splitting. In addition, a sentence simplification component based on a few regular expressions removes unimportant components of a sentence (e.g., *As a matter of fact,*). This processing step does not involve any syntactic parsing. The Question sentiment and target analyzer determines the polarity of the question and identifies the target of a question. For the current experiment, the polarity was determined manually and the target was supplied by us, too.

Given the polarity and the target, the filters we apply identify sentences that are about the target and furthermore extract sentences with positive and negative polarity.

3.1.1 *Preprocessing*

We modified FastSum in order to process blogs by (a) utilizing an HTML parser to extract only text from the blog entries and (b) ignoring boilerplate language in the blogs (e.g., *Response by*). We used the Jericho htmlParser [2] for parsing the HTML documents. Deleting boilerplate language was achieved by a simple filter that (a) computed the density of capitalized words in a sentence[3] and (b) by matching a regular expression that contains frequently used language in blog entries.

3.1.2 *Question sentiment and target analyzer*

[2] http://jerichohtml.sourceforge.net/doc/index.html

[3] Sentences that contain more than 50% capitalized words were automatically excluded.

	Our TAC 2008 Entry	This Research
Target	Supplied by NIST Named entity (manual)	Supplied by us Noun phrase (manual)
Question Analysis	Regular patterns and keywords	none

Table 1: Differences between our TAC 2008 system and the setting presented here.

Unlike in our TAC system, we did not use question analysis (Table 1); targets were noun phrases rather than entities, and these were manually provided by the authors.

3.1.3 Filtering

As an initial filter, we ignore all sentences that do not have at least two exact word matches or at least three fuzzy matches with the topic description.[4] Sentences were selected according to their sentiment and whether they were related to the target of the questions.

3.1.3.1 Sentiment tagging.

We implemented a sentiment polarity tagger largely based on unigram term lookup. While ultimately we believe that polarity tagging cannot be reduced to a context-insensitive word lookup task, we wanted to experiment with the impact of gazetteers, since such simplistic methods represent the largest part of the published literature, and they provide a baseline that more complex methods should benchmarked against.[5] We created gazetteers of positive and negative polarity-indicating terms based on the *General Inquirer* [26], extended it, and also eliminated some erroneous entries. We then proceeded to incorporate morphological variations of the words already in the gazetteer to improve coverage, and manually eliminated errors created in this process.[6]

We noticed that many gazetteer entries were ambiguous in their status, namely, whether sentiment polarity-bearing and *not* polarity-bearing. We thus decided to build two sets of polarity gazetteers, one as described above, and another one based on appreciable manual pruning, where all potentially ambivalent entries were eliminated to improve precision. For example, the entry *incompetent* was not pruned because it always expresses a negative sentiment, whereas *dependent* was removed from the second gazetteer since it may or may not be used in a neutral sense, depending on context.

The tagging itself was based on looking up tokens, counting positive and negative instances, and assigning a label as follows:

$$\left.\begin{array}{lll} \texttt{NEGATIVE} & if & polarity < -1 \\ \texttt{NEUTRAL} & if & -1 \leq polarity \leq 1 \\ \texttt{POSITIVE} & if & polarity > +1 \end{array}\right\} where$$

$$polarity = (\#PositiveTok - \#NegativeTok)/\#AllTok.$$

[4]Fuzzy matches are defined by the OVERLAP similarity [2] of at least 0.1.

[5] We also implemented a simple negation detection; however, this was not used as it did not outperform a system without negation detection.

[6]This was not technically necessary, since non-words will almost never be looked up.

Polarity	Precision	Recall	F-score (F1)
Positive	46.97%	51.67%	49.21%
Negative	59.52%	*33.78%*	43.10%
Neutral	61.99%	73.10%	67.09%
Overall	58.10%	58.10%	**58.10%**

Table 2: Component evaluation of sentiment polarity tagger.

We developed a test set of 528 sentences randomly extracted from the **BLOG06** subset [17] that was provided by NIST as development data for the TAC 2008 sentiment summarization pilot task [9]. Each sentence (segmented automatically by running the pre-processing pipeline of our system) was hand-labeled by a human reviewer with one of the three labels `NEGATIVE`, `NEUTRAL`, or `POSITIVE`. We performed a component-based evaluation specifically on the sentiment polarity tagger. Since it is based on gazetteers, it does not require a training corpus. The results, given in Table 2, make it clear that recall issues for the negative class are a weak link of this approach. By contrast, the overall F1 score for the method was 58.1%.

3.1.3.2 Target matching.

In the FastSum system used for opinion summarization, we also employed a technique that determined whether the sentence contains the target entity from the query. Target matching was also used for the current experiments, even though a target description was more abstract than the target definition in the TAC competition (e.g. *Whole Foods* vs. *gender discrimination*).[7]

Note that the target need not explicitly be present in the sentence under consideration as long as it is present in a decaying window centered on a target description. We matched words with the target via the Jaro Winkler similarity function in order to account for misspellings of names (e.g., *Mayor Guiliani*). We used the Cosine window function for assigning "targetness" scores to words following an identified target. Hence, a subsequent sentence may still be considered for inclusion in the summary, because the sentence is close to a target description in the previous sentence. Future work needs to focus on how sentences can be identified that are concerned with the legal question at hand and how they can be separated from sentences that are not relevant.[8]

3.2 Features for the SVM sentence ranker

Features are mainly based on frequencies of words in sentences, clusters, documents and topics. The features we used can be divided into two sets: word-based and sentence-based. Word-based features are computed based on the relative frequency of words for different segments (i.e., cluster, document, topic title and description). At runtime, the different relative frequencies of all words in a candidate sentence, s, are added up and normalized by the length $|s|$. Sentence-based features include the length and position of the sentence in the document.

Topic title frequency: the relative topic title word frequency

[7]The targets for the legal questions we used are listed in Table 8.

[8]For more details on how the target analyzer was implemented, see [24].

Scope	Engine	Properties (selected)
General Blog Search Engines (Focus: *Blogosphere*)	technorati.com	Includes authority score
	blogsearch.google.com	Date or relevance ranking
	www.blogsearchengine.com	Focus on higher quality content
Legal Blog Search Engines (Focus: *Blawgosphere*)	www.blawg.com	Generally shorter entries
	blawgsearch.justia.com	Date or relevance ranking
	www.blawgrepublic.com	Generally shorter entries

Table 3: List of Blog Search Engines and Their Properties.

for a title $\mathcal{T}$ given a sentence s: $\frac{\sum_{i=1}^{|s|} f_{\mathcal{T}}(t_i)}{|s|}$, where $f_{\mathcal{T}} = \begin{cases} 1 & : \ t_i \in \mathcal{T} \\ 0 & : \ otherwise \end{cases}$

Topic description frequency: the relative topic description word frequency for a description $\mathcal{D}$ given a sentence s: $\frac{\sum_{i=1}^{|s|} f_{\mathcal{D}}(t_i)}{|s|}$, where $f_{\mathcal{D}} = \begin{cases} 1 & : \ t_i \in \mathcal{D} \\ 0 & : \ otherwise \end{cases}$

Content word frequency: the relative content word frequency $p_c(t_i)$ of all content words $t_{1..|s|}$ occurring in a sentence s. The content word probability is defined as $p_c(t_i) = \frac{n}{N}$, where n is the number of times the word occurred in the cluster and N is the total number of words in the cluster: $\frac{\sum_{i=1}^{|s|} p_c(t_i)}{|s|}$

Document frequency: the relative document frequency $p_d(t_i)$ of all content words $t_{1..|s|}$ occurring in a sentence s. The document probability is defined as $p_d(t_i) = \frac{d}{D}$, where d is the number of *documents* the word t_i occurred in for a given cluster and D is the total number of documents in the cluster: $\frac{\sum_{i=1}^{|s|} p_d(t_i)}{|s|}$

Headline frequency: the relative headline word frequency of all content words in a sentence s. The headline probability is defined as $p_h(t_i) = \frac{h}{H}$ where h is the number of times the word occurred in the headline and H is the total number of words in the headline: $\frac{\sum_{i=1}^{|s|} p_h(t_i)}{|s|}$

Sentence length: a binary feature with a value of 1 if the number of words is between 8 and 50 and zero otherwise.

Sentence position (binary): indicates whether the position of the sentence is less than a predefined threshold.

Sentence position (real): the ratio of the sentence position over the number of sentences in the document.

3.3 Training the sentence ranking

In order to learn the feature weights, we trained a regression SVM [13] on Document Understanding Conference (DUC) 2007 [8] news data using the same feature set. In regression, the task is to estimate the functional dependence of a dependent variable on a set of independent variables. In our case, the goal is to estimate the "summary-worthiness" of a sentence based on the given feature set. In order to get training data, we computed the word overlap between the sentences from the document clusters and the sentences in DUC model summaries. We associated the word overlap score to the corresponding sentence to generate the regression data. Note that this is the overlap score based on exact matches, and not the OVERLAP score used for computing fuzzy matches, as described in the previous section.

3.4 Redundancy removal

As a last step, we use the pivoted QR decomposition to handle redundancy [7]. The basic idea is to avoid redundancy by changing the relative importance of the rest of the sentences based on the currently selected sentence. The final summary is created from the ranked sentence list after the redundancy removal step.

4. DATA

We initially examined several of our query logs for TAC 2008-like queries. These included queries to our law review and legal journal databases. Although there were queries of a sentimental nature to be found (on court "dispositions" and "attitudes" toward certain subjects), we believed they did not exist in sufficient quantities to claim them as representative. As an alternative, we subsequently began to investigate TAC-inspired queries run against some of the currently prominent blog search engines.

In order to harness relevant and substantive blog entries on legal topics of interest, we examined six blog search engines. These included three which focus specifically on the *blawgosphere* and three that focus generally on the *blogosphere*. The engines investigated are presented in Table 3.

We were able to make two significant observations concerning these engines. First, the entries tended to be returned in reverse chronological order, though Justia and Google gave users the option of either date-ranked (current first) or relevance-ranked. And, second, the entries these engines returned were short, with an average length of just a couple of paragraphs. The one notable exception to the first observation was technorati.com, which provides an *authority score* assigned to entries that corresponds to in-links (cites) to that blog.[9] The exception to the second observation was blogsearchengine.com, whose entries tended to have an average length exceeding a page. blogsearchengine.com claims to focus on higher quality blogs, and this may be responsible for its longer search results.[10]

Given that our objective was to obtain both informative and substantial amounts of content as input to our summarization engine, following our trials, we eventually decided to

[9]Although this is a useful metric, it also holds a bias against newer and potentially valuable postings. This problem and its solutions are explored in [5].

[10]blogsearchengine.com uses icerocket.com as its core engine. Given that it is an established engine whose results support its claim influenced our decision to harness it as our source of input documents.

use blogsearchengine.com as our core blog search engine. Although it is a general blog search engine rather than one focusing on legal blogs, because of the quality and thoroughness of its entries, in a *de facto* manner, its results were at least as relevant in a legal sense as the results from those claiming to focus exclusively on the blawgosphere. Moreover, in some respects these divisions are artificial, insofar as one need not be on a site indexed as a legal blog in order to host a discussion about a legal topic.

Against blogsearchengine.com, we ran dozens of TAC-inspired sentiment (or polarity)-based queries on a spectrum of legal subjects, a significant number involving opponents or proponents of a variety of civil rights disputes. For the experiment reported here, a dozen of these queries were used, with roughly the same number conveying a positive polarity as a negative polarity, as well as ones which were polarity-neutral though asking a polarity-laden question, and, finally, one which asked that the two sides of an issue be compared.

5. METHODOLOGY

A workflow diagram for our legal blog summarization process is presented in Figure 3. Several preprocessing steps take place before Web-based blog entries are introduced to the FastSum engine. These include translating the original legal opinion topics into queries and identifying any target entities or concepts within those queries, running the queries through the blog search engine and aggregating the top-ranked results, and passing those results through a "marginal relevance filter" in order to ensure that the entries serving as FastSum input data surpass a minimum relevance criterion. Other input stages correspond to those described in Section 3 and illustrated in Figure 2 (i.e., the query & target analyzer and the feature entry set). FastSum produces a 250 word summary as output, suitable for review by our legal experts. A sample summary is presented in Table 6, while a complete set of the queries and targets used to produce the summaries may be seen in Table 8.

Once blogsearchengine.com was identified as the most effective search engine for our task, we selected twelve queries in question form from our pool, each modeled on TAC opinion summarization queries [9]. They tended to cover a broad spectrum of topics ranging from civil rights to Internet privacy, and from government involvement in the current financial crisis to Darfur. These queries served to generate result sets which were fed into our FastSum system. Two of these queries and their results were used for a training and standardization phase with our reviewers. The remaining ten queries and their results were used to produce actual test summaries for grading.

The top-ten results were gathered for each run, though not all ten-entries were automatically considered.[11] In TAC-like manner, only those entries that were at least marginally relevant to the query were considered, and these were identified using either semi-automatic or manual means. If in the rare event the top-ten results contained zero to one relevant entry only for the given query, the next set of 10 results was considered. This happened only once in our experimental setting.

We used FastSum to produce summaries with a length setting of 250 words. This corresponds to the TAC standard. The pruned version of our gazetteer was used for polarity tagging. FastSum offers summary length as one of its input parameters, yet there are clear tradeoffs in increasing or decreasing this variable. Although additional sentences possess the promise of additional recall of answer-specific material, they also carry with them the risk of less relevant or lower quality information. Given the linguistic quality metric described below, we also ran FastSum with its "remove redundancy" option set.

[11]We used the Lynx program to extract the URLs and Perl's LWT module to download the results in HTML format.

Grade	Meaning	Interpretation
(5)	Very good	On point relative to question, including polarity
(4)	Good	Addresses question, including at least partially the polarity
(3)	Adequate	Marginally relevant to question, independent of polarity
(2)	Poor	May have overlap with question topic, and its polarity
(1)	Very poor	Misses the general point of question, polarity aside

Table 4: Reviewer Guidelines for *Responsiveness*.

Dimensions	Essential Considerations
Grammaticality	no datelines, system internal formatting, fragments, omissions, cap. errors, etc.
Non-redundancy	no unnec. repetition, especially among complete sentences, facts, noun phrases
Referential Clarity	easily identifiable pronouns and noun phrases, same with role in summary
Focus	should have clear focus, sentences' info. should relate only to rest of summary
Structure and Coherence	should be well-structured and organized, sentences tied together, not an info. heap

Table 5: Reviewer Guidelines for *Linguistic Quality*.

Grading of the FastSum summaries was performed by two professional assessors. Both assessors were attorneys with several years of annotation and evaluation experience. To assess the quality of the summaries, we tracked two distinct metrics, (1) *responsiveness* (degree to which information content in solution is relevant to the query), and (2) *linguistic quality*. Both of these metrics were used in the TAC 2008 opinion summarization task [9]. An abbreviated version of the guidelines provided to the reviewers can be viewed in Tables 4 and 5.

One of the key deficiencies of TAC 2008's Opinion Summarization Pilot Task, was that although the organizers established the infrastructure to monitor opinion sentiment, it was not actually taken into consideration in assessing the quality of the resulting summaries. By contrast, we have intentionally and explicitly included polarity performance in our reviewer guidelines, thus making our evaluation environment both more rigorous and practical.

In order to standardize the grading process and ensure that the reviewers were interpreting the grading guidelines consistently, we introduced a training round in which, after reviewing the guidelines with one of the authors, they graded two sample summaries. A grading review summit was held afterwards to discuss the reviewers' application of the guidelines to actual summaries. Differences in reviewer grading were the focal point, in order to produce identical grades between the two reviewers were similar texts encountered in the future.

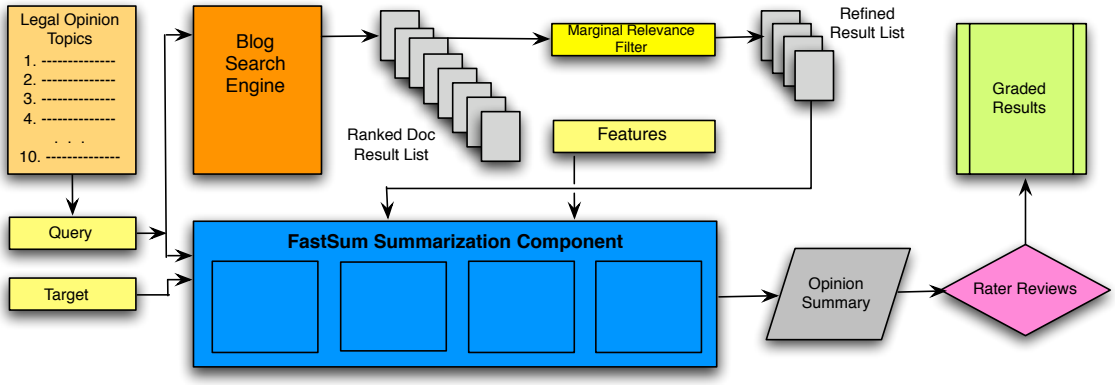

Figure 3: Legal Blog Opinion Summarization Workflow Diagram.

6. RESULTS & DISCUSSION

It is necessary to state that given the limited scope of this study, we make no bold claims about the conclusions that can be drawn. Our findings are, nonetheless suggestive and instructive of opportunities that may exist in the summarization space for robust and tunable systems like FastSum—they may also provide some guidance in the development of more advanced and versatile opinion summarization systems.

A sample result set as presented to the reviewers can be seen in Table 6. It includes (a) topic, (b) question, and (c) summary. The results from our summarization experiment can be seen in Table 7. We note that on average, roughly just under 5 of the 10 blogs returned were found to be relevant to the query and thus eligible to participate in our summarization process. It is difficult to compare this finding with those from TAC, since TAC had already performed the filtering up front and delivered only pre-selected relevant documents at the outset.

One characteristic that distinguishes legal blog entries from more general news documents is their heterogeneity. They are not clean, noise-free, homogeneous documents, but much more diverse, loosely structured, and multi-threaded texts. Such as blogs are, they contain more self-referential text (*"This entry was posted on ...to ... at ..."*). As a result, models that are based on the assumption that continuity exists from one sentence to the next may be ill-conceived. Our current system actually exploits two filters in order to capture and exclude boilerplate language often found in the blogosphere (e.g., *User A, at Time B, on Date C, has just entered this space ...*). Being able to reliably exclude such content from processing workflows is essential for the success of a long term solution.

It is also worth mentioning that the lexical and semantic polarity of our queries were roughly equally divided (5 positive vs. 5 negative questions) along with one neutral and one comparative question.[12] There were slightly higher scores for the summaries of positive polarity topics than negative ones, but part of the reason for this may be explained by their occurrence in the training round, before the reviewers actually had the chance to come together and discuss in detail why they gave the specific grades they did (see Table 7).

In terms of inter-rater agreement, we observe that for six out of the ten test queries, the grades from rater A and rater B were identical, both for *responsiveness* and for *linguistic quality*. For those four test queries where the raters differed, in three of the four cases, they differed by 1 in the *linguistic quality* category. Although there is an appreciable degree of subjectivity involved in both the *responsiveness* and *linguistic quality* judgments, one can see that there is arguably more subjectivity involved in the *linguistic quality* judgment than there is for *responsiveness*, given the reviewer guidelines presented in Tables 4 and 5. Our inter-rater differences concur with this assertion. When calculating a specific Kappa statistic to determine the inter-rater agreement [4], we find a free-marginal Kappa of 0.75, where convention suggests that a Kappa of 0.70 or above indicates adequate inter-rater agreement.[13] Furthermore, one can see from Table 7 that for 80% of their judgments (16 out of 20), the raters concurred, and for the four remaining judgments, they differed by 1 on a five-point scale. In general, with modest training and feedback, the raters were able to establish themselves as reliably consistent graders of basic polarity-based summaries. Also worth noting is that for the most complex question, one that contained two topics (#10), the system performed weakest of all.

Given the consistency of the grades provided by the raters, the average scores for these summaries are within the general range of grades assigned systems participating in TAC's opinion summarization pilot task. So how are scores in the range of 2 to 3 to be interpreted? Given that a score of 5 is perceived to correspond to a human-quality summary, and a 4 to a weaker human quality summary, baseline scores between 2 and 3 may be a reasonable accomplishment at this early point in the development of systems that are addressing multi-document, question and polarity-based summaries.

[12]Neutral query example: *"What is the attitude of the federal courts towards ... prejudice against older Americans?"* Comparative query example: *"Do most Europeans value human rights over expanded commerce and profits? "*

[13]Two versions of Kappa are fixed-marginal multi-rater *Kappa* [4] and a multi-rater variation known as free-marginal *Kappa* [21]. Brennan and Prediger suggest using free-marginal *Kappa* when raters are not forced to assign a certain number of cases to each category (our case) and fixed-marginal *Kappa* when they are.[3]

Topic	*Federal Bailout of Banks and Financial Institutions*
Question	Do most people oppose the federal government bailout of U.S. banks and financial institutions?
Summary	On Tuesday, Mr. Bernanke publicly made the case that one of the most unpopular and most scorned programs in Washington the $700 billion bailout program needs to pour hundreds of billions more into the very banks and financial institutions that already received federal money and caused much of the credit crisis in the first place. Pay attention to the fact the banks got a huge ($350 billion dollars) taxpayer bailout last fall. .... Since last September, no major banks have failed and the credit markets have thawed somewhat. Mr. Obama seems to know it; a week before his swearing-in, he is lobbying Congress to release the other half of the financial industry bailout fund. Yet, they continue, and you pay them. The cost of Obamas two-year program is expected to be around $800 billion. Banks have received $200 billion in fresh capital from the Treasury since last fall and have borrowed hundreds of billions of dollars more from the Fed. Once markets stabilize, the banks will buy their stock back from the government. The most glaring example that the banking system needs even more help is Citigroup. FERRE: Even though banks and financial institutions received 350 billion taxpayer dollars in emergency financial aid, for consumers like Baltiera, the cost of credit is still high. A third is to help set up" bad banks," which would purchase bad assets from financial institutions in exchange for cash or equity in the bad bank. *Your comment has not yet been posted.*

Table 6: Example Output of a Produced Legal Sentiment Summary (cf: item 7 in Table 7).

Such direct comparisons with TAC system performance may currently be unwarranted, however, since TAC uses a different document set (news) in a different domain (general, non-legal) and did not take the polarity of the answers into consideration as was required of our raters. An observation one can make after such a preliminary study is that the problem is a hard one, with many factors contributing to the sub-optimal performance witnessed in many summarization systems. As a result, the baselines remain low, and there continue to exist opportunities for research to make significant contributions to the problem through formal, well-structured experimentation.

7. A PROPOSAL

The annual Text Analytics Conference (TAC) organized since 2008 by the U.S. National Institute of Standards and Technology (NIST) offers research groups a shared task in multi-document summarization, including sentiment summarization [9]. We suggest that it would be valuable to have such shared tasks focusing on the legal domain. Moreover, we propose to the IAAIL community that sponsors ICAIL[14] and to NIST to investigate how much interest there may be in collaborating in such an event. A well-designed, rigorous, and formally evaluated "competition" of this kind could help raise the current baseline performance for question-answering and summarization systems deployed in the legal space. Given the sub-domain, one would expect a great deal of interest from academic and industrial groups as well as government departments and agencies. Such a structured research track could benefit from the collective insights of a community, not unlike other recent and novel TREC tracks, such as TREC Legal, which addresses E-Discovery retrieval issues [18].

[14]International Association of Artificial Intelligence and Law (IAAIL), www.iaail.org

8. CONCLUSIONS

As news and information channels become increasingly congested, especially for professionals such as analysts and lawyers, systems that can identify trends in news streams or construct summaries of sets of documents will become more and more important. If those summaries not only address what the documents are saying but also what their opinion on their subject is, they may be indispensable. The recently held Text Analysis Conference was the first to investigate question answering systems and opinion summarization concurrently. Its evaluation of the latter task, however, failed to explicitly assess the correctness of the polarity of its summary answers.

In conducting a preliminary study relying upon a robust summarization system and expert reviewers, we have shown that using an established multi-document summarization system trained on the news domain, yet redeployed in the legal space, can produce useful and promising summaries of a specific polarity. Although some functionality suitable for homogeneous news documents may be less suitable for heterogeneous legal texts like legal blog entries, as an out-of-the-box implementation, it is encouraging to observe that a first version performs comparable to other systems that have participated in TAC.

The key contributions of this paper are three-fold:

- To our knowledge, it is the first work to perform multi-document opinion-based summarization on postings to the legal blogosphere.
- It extends the TAC evaluation of the opinion summarization task to actually assess the accuracy of the results based on how they satisfy the *polarity* of the question.
- It offers a proposal to the research community, that of IAAIL and NIST, to consider a more formal track

No.	Topic	Polarity	Blogs per Summary	Responsiveness		Linguistic Quality	
				Rater A	Rater B	Rater A	Rater B
Pre-1	Anonymous query logs	+	6	3	3	3	3
Pre-2	Google net neutrality	+	3	2	3	3	2
	Average:		4.5	2.5	3.0	3.0	2.5
1.	Abortion rights	+	5	3	2	2	2
2.	Gay rights	−	7	2	2	2	2
3.	Racial discrimination	−	2	2	2	2	2
4.	Gender discrimination	−	5	1	1	2	2
5.	Age discrimination	neutral	6	2	2	2	3
6.	Strict immigration	+	2	3	3	3	3
7.	Financial bailout	−	6	3	3	2	3
8.	Internet privacy	+	4	2	2	2	2
9.	UN, Dafur violence	−	5	2	2	2	2
10.	Euro. values, Human rights	comparative	2	1	1	1	2
	Average:		4.4	2.1	2.0	2.0	2.3

Table 7: Evaluation Results: Assessment of the Summaries by Human Expert Raters.

to pursue this topic in a more structured, coordinated and in-depth manner.

9. FUTURE WORK

Following our initial experimentation, we envision pursuing a set of instructive extensions to the work presented here.

1. If a set of handcrafted model summaries for each topic had been available, the *nugget pyramid* evaluation method could have been applied to the output summaries [16]. A question that warrants investigation is whether it could be beneficial to incorporate a special treatment of legal language in the evaluation procedure, since in professional blawgs, technical (legal) language is sometimes used and situations can be paraphrased in non-technical language due to the informality prevalent in blogs.

2. A comparison with other summarization systems and techniques would be interesting for the legal blogosphere. We proposed a "shared task" style evaluation in this paper.

3. The machine learning based parts of the FastSum system could benefit from training on various types of blog entries (as opposed to news).

4. The impact that various length *input* result sets have on the final summary and just how pure and noise-free they need to be deserves further study. Furthermore, a setting for the *output* length parameter could be established that leads to an optimal balance between responsiveness and linguistic quality.

5. Summaries could incorporate more structure, and the selection of both the quality (which skeleton to base a summary on) and the quantity (how long should each part be) could take into account the ratio of positive, negative and neutral evidence in the blawgosphere.

6. Most significantly, the core FastSum system could be extended with features or filters specific to the legal domain. We experimented with a generic summarizer for news and blogs, which could be tuned to the legal domain in several ways. For example, a legal dictionary could be incorporated so as to rank sentences higher (i.e., raising their chance of being in a summary) that contain the mention of significant legal concepts.

10. ACKNOWLEDGEMENTS

The authors wish to thank their summary reviewers, Brian Liefeld and Kevin Lane, for the professional and dedicated manner with which they approached the reviewing task. We also appreciate the support of this research provided by Khalid Al-Kofahi, Peter Jackson and James Powell. And lastly, we thank our anonymous reviewers, who helped us improve the quality of this paper.

11. REFERENCES

[1] Kevin D. Ashley and Vincent Aleven. Toward an intelligent tutoring system for teaching law students to argue with cases. In *Proceedings of the Third International Conference on Artificial Intelligence and Law (ICAIL 1991)*, pages 42–52, New York, NY, 1991. ACM.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. In *Proceedings of 16th International World Wide Web Conference (WWW 2007)*, pages 757–766, Banff, Canada, 2007.

[3] Robert L. Brennan and Dale J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3):687–699, 1981.

[4] Jean Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[5] Jack G. Conrad, Jochen L. Leidner, and Frank Schilder. Professional credibility: Authority on the Web. In *Proceedings of the Second Workshop on Information Credibility (WICOW 2008)*, pages 85–88, New York, NY, 2008. ACM.

[6] Jack G. Conrad and Frank Schilder. Opinion mining in legal blogs. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL 2007)*, pages 231–236, New York, NY, 2007. ACM.

[7] John M. Conroy and Dianne P. O'Leary. Text summarization via hidden markov models. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 406–407, New York, NY, USA, 2001. ACM.

No.	Query	Target
(1)	What are the main arguments of pro-choice supporters of abortion rights?	Pro-Choice Arguments Abortion Rights
(2)	What are main reasons why people are anti-gay or oppose gay rights or marriage?	Anti-Gay Arguments by Opponents of Gay Marriage
(3)	How strongly does U.S. law oppose racial prejudice, discrimination, bigotry?	Legal Opposition to Racial Discrimination
(4)	Where is gender discrimination or sex discrimination prosecuted most strongly?	Prosecution of Gender Discrimination
(5)	How have federal courts responded to age discrimination, prejudice against older workers, in the workplace?	Workplace Age Discrimination
(6)	What are arguments of supporters of strict immigration policies?	Strict Immigration Policy Support
(7)	Do most people oppose the federal government bailout of U.S. banks and financial institutions?	Federal Bailout of Banks and Financial Institutions
(8)	How important is Internet privacy to Americans?	Americans Threats Internet Privacy
(9)	Do Internet users support making Web query logs anonymous at some point?	Anonymous Internet Query Logs
(10)	Has Google been a consistent supporter of Net neutrality?	Google's Net Neutrality
(11)	Why does the United Nations actively oppose the violence and so-called genocide in the conflict in Darfur?	UN Opposition to the Violence and Genocide in Darfur
(12)	Do most Europeans value human rights over expanded commerce and profits?	European Values of Human Rights over Commerce

Table 8: Opinion-based Queries and Targets.

[8] Hoa Trang Dang, editor. *Document Understanding Conference (DUC 2007)*. NIST, http://duc.nist.gov/pubs.html#2007, 2007.

[9] Hoa Trang Dang. Update summarization task and opinion summarization pilot task. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, Gaithersburg, MD, Nov. 2008. National Institute of Standards and Technology.

[10] Atefeh Farzindar and Guy Lapalme. Legal text summarization by exploration of the thematic structure and argumentative roles. In *Proceedings of the Workshop On Text Summarization Branches Out*, 2004.

[11] Ben Hachey and Claire Grover. Extractive summarization of legal texts. *Artificial Intelligence and Law*, 14(4):305–345, 2006.

[12] Daniel E. Ho and Kevin M. Quinn. Measuring explicit political positions of media. *Quarterly Journal of Political Science*, 3, 2008.

[13] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pages 133–142. ACM, 2002.

[14] Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: Evaluating and learning user preferences. In *Proceedings of the European Association for Computational Linguistics (EACL 2009)*, Athens, Greece, 2009. ACL.

[15] S. Li, Y. Ouyang, W. Wang, and B. Sun. Multi-document summarization using support vector regression. In *Proceedings of the Document Understanding Conference (DUC 2007)*, 2007.

[16] Jimmy Lin and Dina Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2006)*, pages 383–390, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[17] Craig Macdonald and Iadh Ounis. The TREC Blog06 collection: Creating and analysing a blog test collection. Technical Report DCS Technical Report TR-2006-224., University of Glasgow, Glasgow, Scotland, UK, 2006.

[18] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. TREC-2008 legal track overview. In *The Seventeenth Text REtrieval Conference (TREC 2008), Proceedings*, Gaithersburg, MD, Nov. 2008. National Institute of Standards and Technology (NIST).

[19] Paul Over, Hoa Dang, and Donna Harman. DUC in context. *Information Processing and Management*, 43(6):1506–1520, 2007.

[20] Paulo Quaresma and Irene Pimenta Rodrigues. A question-answering system for Portuguese juridical documents. In *Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL 2005)*, pages 256–257. ACM, 2005.

[21] Justus J. Randolph. Free-marginal mutirater kappa: An alternative to fixed-marginal multirater kappa. In *Joensuu Learning and Instruction Symposium*, Joensuu, Finland, Oct. 2005. ERIC Document Reproduction Service No. ED490661.

[22] M. Saravanan, B. Ravindran, and S. Raman. *Proceedings of the 19th Annual Conference of Legal Knowledge and Informa- tion Systems (JURIX 2006)*, chapter Improving Legal Document Summarization Using Graphical Models, pages 51–60. IOS Press, 2006.

[23] Frank Schilder and Ravikumar Kondadadi. FastSum: Fast and accurate query-based multi-document summarization. In *Proceedings of the Joint Annual Meeting of the Association for Computational Linguistics and the Human Lanugage Technology Conference (ACL-HLT 2008)*, pages 205–208, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[24] Frank Schilder, Ravikumar Kondadadi, Jochen L. Leidner, and Jack G. Conrad. Thomson Reuters at TAC 2008: Aggressive filtering with FastSum for update and opinion summarization. In *Proceedings of the First Text Analysis Conference (TAC 2008)*, pages 396–405, Gaithersburg, MD, 2008. NIST.

[25] Frank Schilder, Jochen L. Leidner, Jack G. Conrad, and Ravikumar Kondadadi. Polarity filtering for sentiment summarization. In *Poster presented at the First Text Analysis Conference (TAC 2008)*, Gaithersburg, MD, 2008. NIST.

[26] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, Cambridge, MA, 1966.

[27] Ellen M. Voorhees and Donna K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press, Cambridge, MA, 2005.