Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
An Adjunct to the Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing

Identifying Relevant Text Fragments to Help Crowdsource Privacy Policy Annotations

Rohan Ramanath, Florian Schaub, Shomir Wilson,
Fei Liu, Norman Sadeh, Noah A. Smith
School of Computer Science
Carnegie Mellon University
{rrohan,fschaub,shomir,feiliu,sadeh,nasmith}@cs.cmu.edu

Abstract

In today's age of big data, websites are collecting an increasingly wide variety of information about their users. The texts of websites' privacy policies, which serve as legal agreements between service providers and users, are often long and difficult to understand. Automated analysis of those texts has the potential to help users better understand the implications of agreeing to such policies. In this work, we present a technique that combines machine learning and crowdsourcing to semi-automatically extract key aspects of website privacy policies that is scalable, fast, and cost-effective.

Overview

Privacy policies are verbose, often complicated legal documents that protect the interests of online service providers. McDonald & Cranor (2008) showed that, if users were to read each privacy policy of websites they access in a year, they would spend an unreasonable fraction of their time doing it, but would often still not be able to answer basic questions about the policies' meaning. Unsurprisingly, many people do not read them (Federal Trade Commission 2012).

It is important for users to understand the implications of agreeing to these policies. There have been various attempts to standardize the format of privacy notices (Cranor 2012), such as P3P or privacy nutrition labels, so that they are easier to understand and allow users to have better control of their personal information. In the past, most qualitative analyses of policies have been carried out on small sets of documents. Building user-oriented tools that help users understand the contents of policies will require a larger scale of analysis. Having legal experts interpret privacy policies is expensive and slow, but one potential alternative is to enlist crowdworkers (Reidenberg et al. 2014). However, the length and complexity or privacy policies make them difficult to interpret in their entirety for the average crowdworker. Prior crowdsourcing research has shown that hard tasks can be reformulated by decomposing them into smaller solvable tasks (Kittur et al. 2011). Crowdsourcing for relevance extraction is a possible candidate for decomposition, which has been demonstrated in the development of information retrieval systems. However, such relevance judgements have been used for evaluation (Alonso and Mizzaro 2012) or other purposes (Parameswaran et al. 2012) and not yet to decompose the task itself. The closest work in the crowdsourcing literature that tackles issues of text data being complex, high-dimensional and ill-structured is that of André et al. (2014). However, they are only interested in classifying the data (via clustering).

Assuming that the task is to crowdsource answers to a list of questions, some naïve methods to decompose the task are apparent: (1) Tasks can be issued that require workers to answer only one question about the given document. Since a given issue can be discussed in multiple locations within the policy, the worker will have to read the entire policy to answer the question. Reading irrelevant parts a policy substantially increases the time required to complete the task. (2) Combinations of every paragraph and every question could be presented in separate HITs. This would increase the number of HITs significantly. If $|P|$ is the number of paragraphs in the policy and $|Q|$ is the number of questions, we would now have $|P| \times |Q|$ tasks per policy. Another inherent problem with this approach is that, if the same question has been answered differently when presented with two different paragraphs, there is no straightforward way to automatically determine the "correct" answer for the policy as a whole. (3) A subdivision (section) of the privacy policy and all questions. Our group tried this approach in a pilot study and found that reconciling different answers from on a set of section-based answers for a question is difficult and fraught with questionable assumptions (Ammar et al. 2012).

In this work, we introduce a novel technique to decompose the crowdsourcing task of answering questions about privacy policies that does not require workers to read entire policies. Our approach learns parts of the policy that are relevant to a given question, allowing us to present each question as a separate task along with those policy segments learned to be "relevant." This solution offers the promise of improved performance on crowdsourcing of policy annotations at a reduced price, and likely without loss of annotation accuracy.

Dataset & Experimental Setup

We use the crowdsourced privacy policy dataset provided by Ramanath et al. (2014). To make the task more user-friendly,

we used an external HIT integrated into the Mechanical Turk workflow. Along with a qualified team of privacy and legal experts we drafted a set of 9 questions that captured information about collection of a user's contact, location, health, and financial information, sharing with third parties, and deletion of data. To establish the performance of a baseline approach, we set up a Mechanical Turk task that presented workers with a full-length privacy policy along with the 9 questions to be answered. We issued tasks for 10 policies, and each was performed by 10 workers who took on average 22 minutes, costing $6 per policy per Turker. The same tasks were completed by lawyers and privacy experts, and their answers were used as the gold standard to evaluate the quality of the crowdsourced annotations. To accumulate training data, we presented a set of 25 policies (including the earlier 10) to students who study privacy. In addition to answering the questions, they were asked to highlight sentences in the policy that helped them answer each question. These highlighted sentences were later processed to extract regular expressions that identify the presences of a particular topics in a policy.

Proposed Solution

We use an ℓ_2-regularized logistic regression classifier that labels a given input paragraph as "relevant" or "not relevant" with respect to an input question. *n-grams* ($n \leq 3$) are used as the basic features after stemming and stopword removal. Additionally, we use the regex extracted earlier to define features on whether the regex matched text chunks in the paragraph and *n-grams* ($n \leq 3$) of the matched text.

Equipped with the resources to identify relevant paragraphs to a given question, we present two models to cost-effectively and quickly annotate the policies without loss of annotation accuracy. In the first, we still present the entire policy with relevant parts highlighted, along with the 9 questions. The hope in this case is that Turkers are able to answer the questions using the highlighted parts along with some context from the rest of the policy. In the second model, the earlier single task with 9 questions is split into 9 sub-tasks, one for each question. For each of these sub-tasks, we show only the paragraphs relevant to the specific question, in the original order of occurrence, rather than the full policy. The rationale is that Turkers should be able to answer the question based on the relevant paragraphs alone. Since significantly less text is presented to the Turker, the task should be faster to complete ($\sim$45s per question) and, hence, cost less. Turkers can further give feedback in case the answer is not in the highlighted region of the policy. We use this information to create better and more refined patterns.

Preliminary analysis using the highlighted sentences showed that the answer to a given question is concentrated in a particular region of the policy as opposed to being scattered all over. Figure 1 shows an example. Further analysis revealed that most policies word the salient clauses similarly. This scenario motivates the use of regular expressions to capture sentences that are likely to contain the answer to a given question. The patterns are designed to hierarchically capture broader topics (such as collection, sharing) first, and these results are subsequently filtered to capture narrower topics (such as specific types of collected information).

collect contact information in `bloomberg.com` *policy:*

> BLP or a designated service provider collects personal information that you voluntarily provide on this Web site, on third-party websites and applications when subscribing to the digital edition of Bloomberg Businessweek, or offline in connection with your Bloomberg Businessweek print edition subscription, which may include your name, address, e-mail address, credit card number, billing information, etc.

Figure 1: When asked the question, *Does the policy state that the website might collect contact information about its users?*, all annotators agreed that the answer was contained in the sentences above.

Outlook

In this paper, we outlined a method to improve policy segment and task matching based on content relevance. We are in the process of refining and evaluating the proposed technique in a series of crowdsourcing experiments. The eventual incorporation of our method into the document annotation and data acquisition process offers the promise of furthering the study of automated analysis of privacy policies.

References

Alonso, O., and Mizzaro, S. 2012. Using crowdsourcing for trec relevance assessment. *Information Processing and Management* 48(6):1053–1066.

Ammar, W.; Wilson, S.; Sadeh, N.; and Smith, N. A. 2012. Automatic categorization of privacy policies: A pilot study. Technical Report CMU-LTI-12-019, Carnegie Mellon Univ.

André, P.; Kittur, A.; and Dow, S. P. 2014. Crowd synthesis: Extracting categories and clusters from complex data. In *Proc. CSCW '14*, 989–998. ACM.

Cranor, L. F. 2012. Necessary but not sufficient: Standard mechanisms for privacy notice and choice. *Journal of Telecommunications and High Technology Law* 10.

Federal Trade Commission. 2012. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers.

Joel R. Reidenberg et al. 2014. Disagreeable privacy policies: Mismatches between meaning and users understanding. In *Proceedings of 42nd Research Conference on Communication, Information and Internet Policy*, TPRC'14.

Kittur, A.; Smus, B.; Khamkar, S.; and Kraut, R. E. 2011. Crowdforge: Crowdsourcing complex work. In *Proc. UIST '11*, 43–52. ACM.

McDonald, A. M., and Cranor, L. F. 2008. The cost of reading privacy policies. *I/S: J Law & Policy Info. Soc.* 4(3).

Parameswaran, A. G.; Garcia-Molina, H.; Park, H.; Polyzotis, N.; Ramesh, A.; and Widom, J. 2012. Crowdscreen: Algorithms for filtering data with humans. In *Proc. SIGMOD '12*, 361–372. ACM.

Ramanath, R.; Liu, F.; Sadeh, N.; and Smith, N. A. 2014. Unsupervised alignment of privacy policies using hidden Markov models. In *Proc. ACL '14*.