Demystifying Privacy Policies with Language Technologies: Progress and Challenges

Shomir Wilson*, Florian Schaub*, Aswarth Dara*, Sushain K. Cherivirala*, Sebastian Zimmeck†, Mads Schaarup Andersen*, Pedro Giovanni Leon‡, Eduard Hovy*, Norman Sadeh*

*Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
shomir@cs.cmu.edu, sadeh@cs.cmu.edu

†Columbia University
116th Street & Broadway
New York, NY 10027, USA

‡Stanford University
450 Serra Mall
Stanford, CA 94305, USA

Abstract

Privacy policies written in natural language are the predominant method that operators of websites and online services use to communicate privacy practices to their users. However, these documents are infrequently read by Internet users, due in part to the length and complexity of the text. These factors also inhibit the efforts of regulators to assess privacy practices or to enforce standards. One proposed approach to improving the status quo is to use a combination of methods from crowdsourcing, natural language processing, and machine learning to extract details from privacy policies and present them in an understandable fashion. We sketch out this vision and describe our ongoing work to bring it to fruition. Further, we discuss challenges associated with bridging the gap between the contents of privacy policy text and website users' abilities to understand those policies. These challenges are motivated by the rich interconnectedness of the problems as well as the broader impact of helping Internet users understand their privacy choices. They could also provide a basis for competitions that use the annotated corpus introduced in this paper.

Keywords: Privacy, Internet, annotation, corpus, crowdsourcing.

1. Introduction

Websites' privacy policies are the common (if not nearly pervasive) mechanism by which website operators inform users how their data will be collected, protected, shared, or otherwise processed. However, studies have shown that the average Internet user reads few of the privacy policies of websites they visit (Federal Trade Commission, 2012), would need a substantial amount of time to do so, and even then would have difficulty understanding what those policies mean (McDonald and Cranor, 2008). Moreover, the length and complexity of these documents are a hindrance to policy regulators who are tasked with assessing and enforcing compliance with legal and regulatory requirements. These problems have led to the assessment that the current "notice and choice" model of online privacy is broken (Reidenberg et al., 2014).

These impractical aspects of privacy policies pose a ripe challenge that several efforts have tried to resolve. Some, most notably P3P (W3C, 2006), have relied on voluntary cooperation from website operators to provide formally-specified data on their privacy practices, which are presented to users or applied to browser settings such as cookie management. However, website operators have been hesitant to supply their privacy policies in a machine-readable format, to the extent that P3P has not been widely adopted. Other efforts, such as Terms of Service; Didn't Read (Roy, 2016), have relied chiefly on volunteers to annotate privacy policies. This approach has limits as well, as it relies solely on the attentiveness and dedication of its community to function.

An opportunity exists for language technologies to provide some degree of automation to "bridge the gap" between privacy policies and their audiences. In some ways this is a familiar problem domain; natural language processing on legal text is an active area of research and the legal community has begun to recognize it as well (Mahler, 2015). However, the salience of the problem to the average Internet user's experience makes it unique among applications of natural language processing to legal text. Its timeliness to rising concerns about digital privacy is also a strong motivating factor.

We report on the progress of the Usable Privacy Policy Project, an ongoing effort to use crowdsourcing, natural language processing, and machine learning to extract key information from privacy policy text and present that information to Internet users and policy regulators. We outline our efforts to enlist human annotators—both crowdworkers and expert readers—to gather information from privacy policy text. We have applied a fine-grained policy annota-

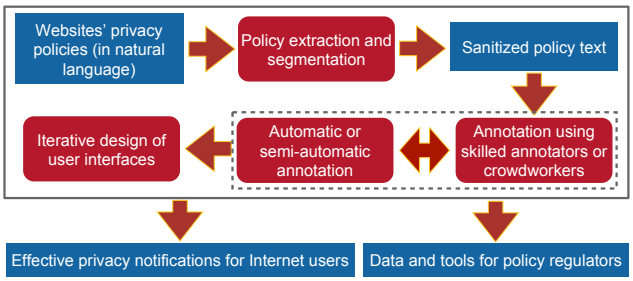

Figure 1: The structure of our approach for processing website privacy policies.

tion procedure to a set of 115 privacy policies (267K words) to label them with a total of 23K data practices, 128K practice attributes, and 103K annotated text spans. This corpus of privacy policies is unprecedented in its size and detail, and we plan to release it to the research community to encourage exploration of this topic. Finally, we describe several research challenges for automating the annotation and analysis of privacy policies. These problems are motivated by the status quo in usable online privacy and the potential broader impacts of improving language technologies.

2. Making Privacy Policies Usable

The Usable Privacy Policy Project[1] (Sadeh et al., 2013) is an NSF-funded Frontier Project with the goals of automatically or semi-automatically extracting key details from privacy policy text, presenting those details to Internet users in formats that respond to their needs, and enabling large-scale privacy policy analysis to inform regulators and key policymakers. We concentrate below on the aspects of the project that are most relevant to language technologies, to motivate the challenges we wish to share with this research community.

Figure 1 illustrates our overall approach to processing privacy policy text. We assume no cooperation on the part of website operators; thus our pipeline begins with website privacy policies as they are generally presented to users, in natural language with some HTML markup. Each privacy policy is downloaded, sanitized to remove extraneous page elements, and then (if appropriate for later stages) divided into segments that roughly correspond to paragraphs. Privacy policies or policy segments are then annotated with information on privacy practices using a combination of human annotators and automatic methods currently under development. We are improving this stage incrementally, by first gathering data from policies with the help of human annotators and then using that data to train models that predict aspects of policy contents. Over time, we expect to increasingly automate the annotation process, limiting the need for human annotators to tasks where human judgement is strictly necessary.

Finally, the results of the annotation process must be presented to Internet users and policy regulators in ways that are responsive to their needs. Our project is developing browser plugins that will show users the privacy practices of websites as they visit them. This is inspired in part by prior work on privacy nutrition labels (Kelley et al., 2009) and privacy profiles (Liu et al., 2014), recognizing the substantial challenges to showing users information about privacy practices. Our team has also developed a data exploration website to showcase the results of some of our annotations. We introduce this site in the next section. The intended outcomes of this project include a transfer of technologies and analysis results to industry, regulators, and policymakers, to ensure a lasting impact of the work.

[1] https://www.usableprivacy.org

3. Collecting Information from Privacy Policies

We have taken a two-pronged approach to human annotation of privacy policy contents, by creating a question-based annotation tool for crowdworkers and a fine-grained annotation tool for expert annotators. We describe both below and explain the future intersection of these approaches.

3.1. Crowdsourcing

Crowdsourcing is a well-recognized method of solving problems that are difficult for computers but easy for humans (Quinn and Bederson, 2011). However, Internet users interpret privacy policies only with great difficulty (Jensen and Potts, 2004; McDonald and Cranor, 2008), potentially ruling out crowdsourcing for this problem. To overcome this limitation, we investigated the accuracy of crowdworkers' answers to questions about privacy policies when multiple workers' answers are aggregated.

Figure 2 shows the interface of our privacy policy annotation tool for crowdworkers. For the task, crowdworkers read the text of a privacy policy, answered nine multiple-choice questions about the privacy practices that it described, and highlighted the text that answered each question. Ten crowdworkers completed this task for each policy, and we experimented with aggregating their answers for each question using a confidence threshold: if the fraction of annotators who chose the most popular answer to a question about a privacy policy was no less than a specified percentage, that answer was designated the crowd's answer. If the most popular answer did not meet the percentage, then it was deemed that the crowd had not produced an answer. Varying the threshold thus permitted tuning for only high agreement answers (high threshold) or for broad coverage (low threshold). Separately, we also experimented with using a logistic regression model to predict which paragraphs were relevant to each question. We highlighted these paragraphs for crowdworkers and measured their impact on workers' accuracy, speed, and confidence.

Our results showed that requiring a high level of agreement produced relatively high accuracy while retaining substantial coverage (Wilson et al., 2016). For example, setting the threshold at 80% agreement resulted in crowd answers for 69% of question-policy pairings and 96% of those answers matched experts' interpretations. We also observed a slight increase in task completion speed with the provision of relevance highlights, without a negative impact on accuracy. Moreover, in an exit survey, annotators who used the interface with highlights expressed greater ease with understanding legal texts than annotators who did

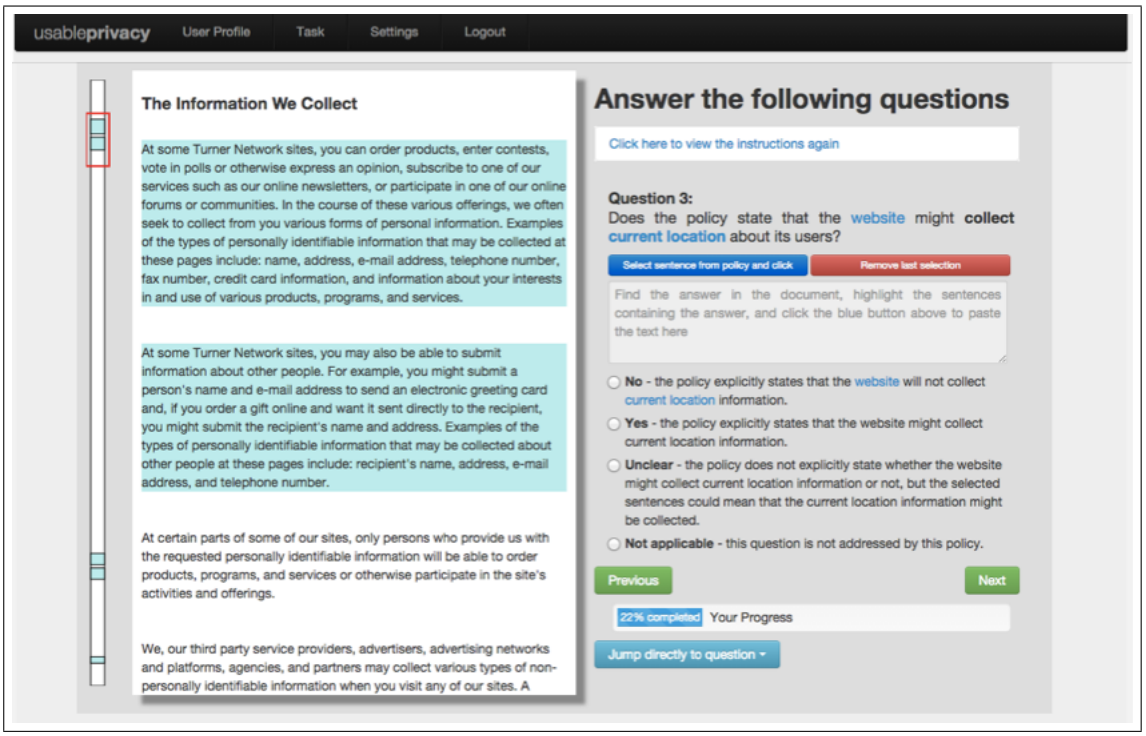

Figure 2: The policy annotation tool for crowdworkers.

not see highlights. This suggests that the task is easier for crowdworkers when highlights are provided, and providing them may reduce fatigue and increase endurance for similar labeling tasks.

3.2. Expert Annotation

The crowdworker annotation tool collected answers to simple questions about privacy policies, and a need remains for annotations that better capture the nuances of privacy practices. Solving this entirely with crowdworkers is highly complex and challenging: even after defining a sufficiently fine-grained annotation scheme, the task is too intricate to give to crowdworkers and interdependencies within it must be discovered to decompose it. Thus, we created a fine-grained annotation tool and enlisted *expert workers* (graduate students in law) to annotate a set of privacy policies. We are planning for an initial release of a corpus of 115 privacy policies plus annotations on the Usable Privacy Policy Project website (URL in Footnote 1). The next steps will be to identify annotation subtasks that can be automated and other subtasks that are suitable for crowdsourcing.

The intent of the fine-grained annotation scheme is to capture the relationship between the data practices (i.e., axioms about how a website user's data is collected, shared, or otherwise applied) intended by a privacy policy and the specific segments of text that express those practices. Each data practice belongs to one of ten categories that broadly express its genus:

1. First Party Collection/Use
2. Third Party Sharing/Collection
3. User Choice/Control
4. User Access, Edit and Deletion
5. Data Retention
6. Data Security
7. Policy Change
8. Do Not Track
9. International and Specific Audiences
10. Other

Each of these ten categories is associated with a set of attributes, and each attribute is associated with a set of potential values. For example, *User Choice/Control* has five attributes: *Choice Type*, *Choice Scope*, *Personal Information Type*, *Purpose*, and *User Choice*. An attribute may be required (the annotator must select a value for the attribute when creating a practice) or optional (since policies are sometimes vague). Crucially, most values are required to be associated with a span of text in the privacy policy.

Figure 3 shows the interface of the fine-grained annotation tool. Expert workers read one policy segment at a time. To create a data practice, the annotator first selects a category and then, for each attribute, specifies a text span (by clicking and dragging) and a value. For example, Figure 3 shows a partly instantiated *User Choice/Control* practice: *Opt-in* is the selected value for *Choice Type*, and it is associated with the highlighted span of text in the policy fragment. A value has been selected for attribute *Purpose* as well, though its text span is not shown. Values have not yet been selected for three attributes, though the lack of an asterisk by attribute *User Type* indicates it is optional.

We have collected annotations for a set of 115 website privacy policies. Websites were selected to represent diverse coverage of sectors (e.g., entertainment, e-commerce,

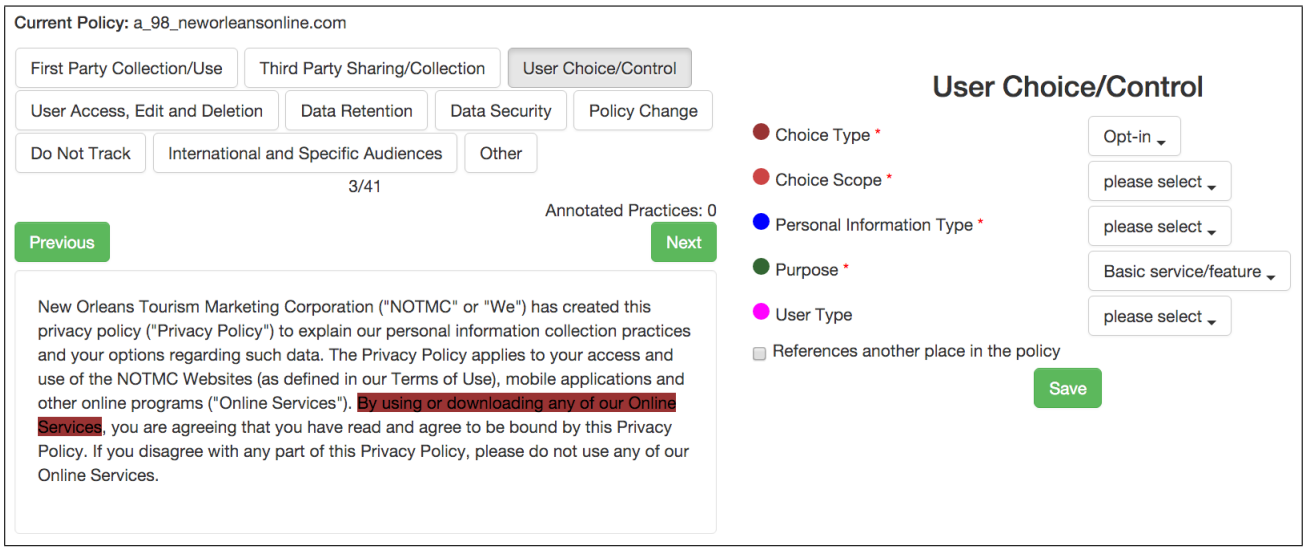

Figure 3: The fine-grained policy annotation tool, for expert annotators.

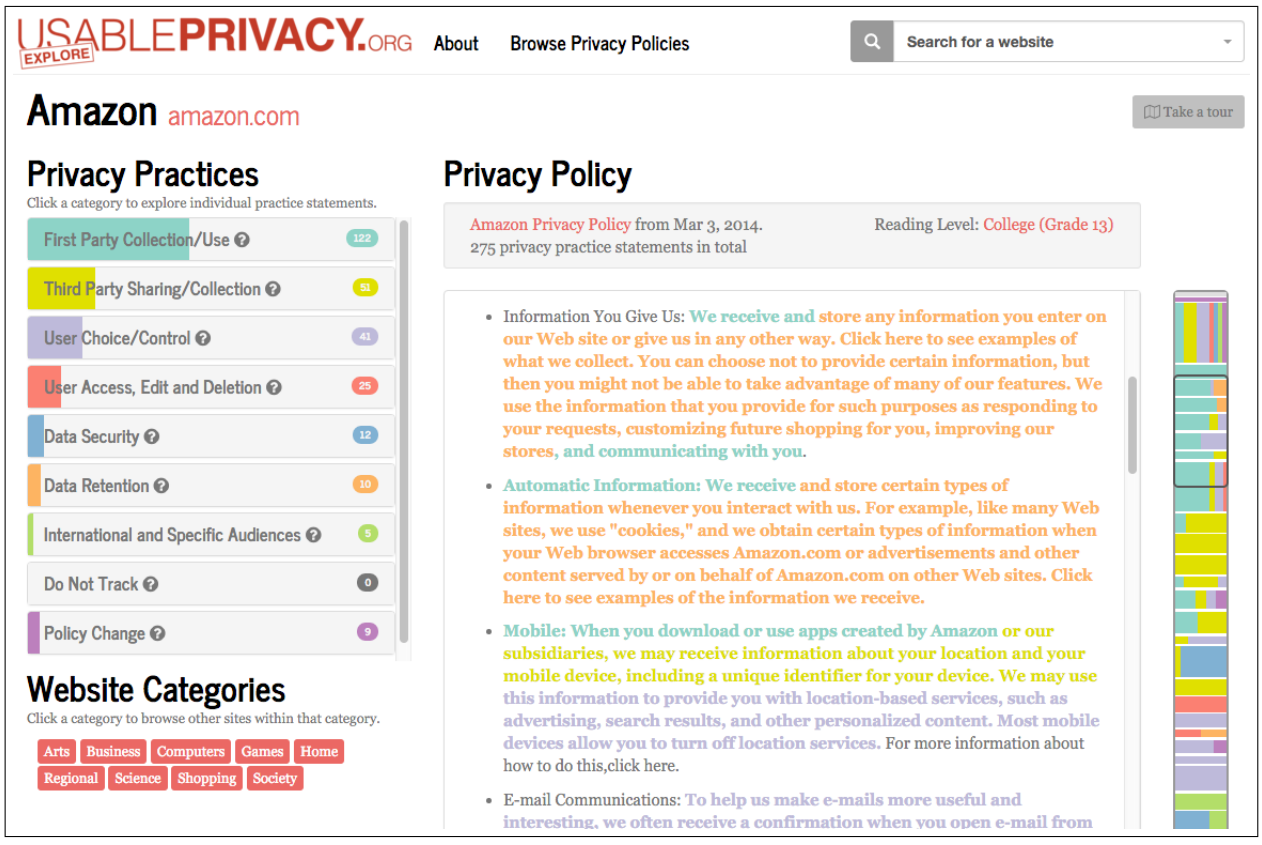

Figure 4: Viewing a sample policy on the data exploration site.

reference, education, etc.) and levels of popularity (as determined by Alexa.com rankings). Each website was annotated by three expert workers who worked independently.

The annotated corpus represents an unprecedented survey of online privacy practices as well as a unique, specialized language resource. In aggregate, the privacy policies consist of 267K words, and the expert workers produced 23K data practices, 128K practice attributes, and 103K annotated text spans associated with those attributes. Notably these counts of annotations represent unconsolidated work—i.e., from three expert workers—and their annotations exhibit some variation. We are exploring several alternatives for identifying and consolidating redundant practices, which is a nontrivial problem because of the complexity of the annotation scheme and thus the variety of modes of divergence.

Finally, Figure 4 is a screenshot of a data exploration

website[2] that we have created to showcase the annotations for the set of 115 website privacy policies. The website allows users to search the collection of privacy policies for specific websites and browse the privacy practices annotated by the expert workers. It also identifies the sectors that a website belongs to, as categorized by DMOZ,[3], and allows the user to compare it to its peers in the same sector.

4. Challenges for the Research Community

The corpus of 115 annotated policies is intended to be a resource for research in natural language processing, usable privacy, and policy analysis. To this end, we challenge these research communities to investigate a family of problems related to the automated analysis of privacy policies. These problems are well-motivated by established topics in natural language processing as well as the difficulties of the "notice and choice" model of online privacy in its current form. Solving them will be progress toward helping Internet users understand how their personal information is used and what choices they can make about that usage. Additionally, policy regulators and creators will have tools to help them monitor compliance with laws and detect trends that require action.

A central challenge of this research direction is the need to annotate privacy policies in a scalable, cost-efficient manner. We have already observed how machine learning can be used to guide human annotators' efforts; for example, the automatically-generated paragraph highlights made the crowdsourcing task (Figure 2) easier for workers. We have also performed preliminary experiments with machine learning to determine that policy segments can be classified automatically for their relevance to practice categories in the fine-grained annotation scheme. These are steps toward a goal of limiting the need for human annotators to small, self-contained tasks that are optimal for crowdsourcing while natural language processing and machine learning take care of the bulk of the analysis. An ambitious (but not completely unreasonable) goal will be to eliminate the need for human annotators altogether. By producing confidence ratings alongside data practice predictions, an automated system could account for its shortcomings by stating which predictions are very likely to be correct and deferring to crowdworkers for predictions that lack firmness.

We propose that the automatic annotation of privacy policies with data practices is a suitable problem for a competitive challenge, and the challenge will advance the state of the art in applicable language technologies. Our corpus of 115 annotated policies contains data that can serve as a gold standard for evaluating solutions. The problem is decomposable into two interrelated subproblems:

- *Prediction of segment relevance to categories*: Given a segment of privacy policy text and one of the ten practice categories, can an automated system predict whether the segment contains a practice belonging to the category? Similarly, can an automated system delimit one or more subsegments of text that correspond with individual data practices?

- *Prediction of values for practice attributes*: Given a segment of policy text and the knowledge that it contains a data practice in a specified category, can an automated system predict the values of the practice's attributes? Similarly, can it identify the text associated with each of those values?

Methods from information retrieval, semantic parsing, coreference resolution, and named entity recognition are particularly relevant to these problems, although a functioning system may use a broad range of methods that we do not immediately anticipate. A sufficiently large space exists for potential solutions that the research community could benefit from an organized challenge as part of a workshop or similar event.

Finally, we are also actively working on or are interested in solving several problems that the corpus will enable us to investigate:

- *Consolidation of annotations from multiple workers*: Under what criteria do a pair of non-identical data practices produced by two annotators refer to the same underlying axiom in the text? Criteria may be observable (i.e., present in the practices' attributes or text spans) or latent (depending on factors such as policy ambiguity or vagueness, which may cause two annotations of an axiom to be divergent without either being in error).

- *Recombination of data practices into a cohesive body of knowledge about a privacy policy*: How do data practices for a privacy policy relate to each other? The answer to this is not contained chiefly in the annotations. For example, two data practices may appear to contradict each other even though they do not, because the reconciliation cannot be represented by the annotation scheme, and thus it is absent from the annotations. Inconsistencies, generalizations, and implications are other examples of potential relationships between data practices. Adding expressiveness to the annotation scheme comes with the tradeoff of greater complexity.

- *Summarization and simplification*: Can the text of a privacy policy be shortened or reworded so that the average Internet user can understand it? A simple test for content equivalence is whether an annotation procedure (by humans or automated methods) produces the same set of data practices for the simplified text and the original text. In practice, Internet users have already demonstrated limited patience with text-based privacy policies, but this problem is nevertheless motivated by the broader goal of making complex texts easier to understand.

- *Optimizing the balance between human and automated methods for privacy policy annotation*: Human annotators and automated annotation both have strengths and weaknesses. The ideal combination in an annotation system will depend on the necessary level of confidence in the annotations and the availability of resources. These resources include human

[2] `https://explore.usableprivacy.org/`

[3] `https://www.dmoz.org/` also used by Alexa.com

annotators, computational power, and training data to create computational models.

- *Identifying sectoral norms and outliers*: Within a sector (e.g., websites for financial services or news), how can we identify typical and atypical practices? A bank website that collects users' health information, for example, would be atypical. When an atypical practice is found, when should it be a cause for concern (or commendation)? Can we recommend websites in a given sector based on an Internet user's expressed privacy preferences?

- *Identifying trends in privacy practices*: The activities that Internet users perform online continue to evolve, and with that evolution the mechanisms for collecting, using, and sharing their data are subject to change. The Internet of Things (IoT) provides a potent example, as sensors collect and share progressively larger amounts of sensitive data. Finding trends in privacy practices will help guide policy regulators to focus their attention on emerging issues.

5. Conclusion

We have described progress toward making privacy policies usable, for the benefit of Internet users, regulators, and policymakers. Additionally, we have presented a rich set of challenges for the research community, engaging efforts in crowdsourcing, natural language processing, and machine learning. The corpus of privacy policy annotations introduced in this paper could also provide a basis for one or more competitions in this area.

6. Acknowledgements

The authors would like to acknowledge the entire Usable Privacy Policy Project team for its dedicated work.

This research has been partially funded by the National Science Foundation under grant agreement CNS-1330596.

7. Bibliographical References

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

Jensen, C. and Potts, C. (2004). Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proc. CHI*. ACM.

Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. (2009). A "nutrition label" for privacy. In *Proc. SOUPS*, pages 4:1–4:12, New York, NY, USA. ACM.

Liu, B., Lin, J., and Sadeh, N. (2014). Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In *Proc. WWW*, pages 201–212, New York, NY, USA. ACM.

Mahler, L. (2015). What is NLP and why should lawyers care? *Law Practice Today*.

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

Quinn, A. J. and Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proc. CHI*, pages 1403–1412, New York, NY, USA. ACM. 00257.

Reidenberg, J. R., Russell, N. C., Callen, A. J., Qasir, S., and Norton, T. B. (2014). Privacy harms and the effectiveness of the notice and choice framework. SSRN Scholarly Paper.

Roy, H. (2016). Terms of Service; Didn't Read. https://tosdr.org/. Accessed 2016-02-15.

Sadeh, N., Acquisti, A., Breaux, T. D., Cranor, L. F., Mcdonald, A. M., Reidenberg, J. R., Smith, N. A., Liu, F., Russell, N. C., Schaub, F., and Wilson, S. (2013). The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Technical Report CMU-ISR-13-119, Carnegie Mellon University.

W3C. (2006). The Platform for Privacy Preferences 1.1 (P3P1.1) Specification. `https://www.w3.org/TR/P3P11/`. Last accessed: Mar. 31, 2016.

Wilson, S., Schaub, F., Ramanath, R., Sadeh, N., Liu, F., Smith, N. A., and Liu, F. (2016). Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proc. WWW*, Montreal, Canada. ACM.