Challenges in Automated Question Answering for Privacy Policies

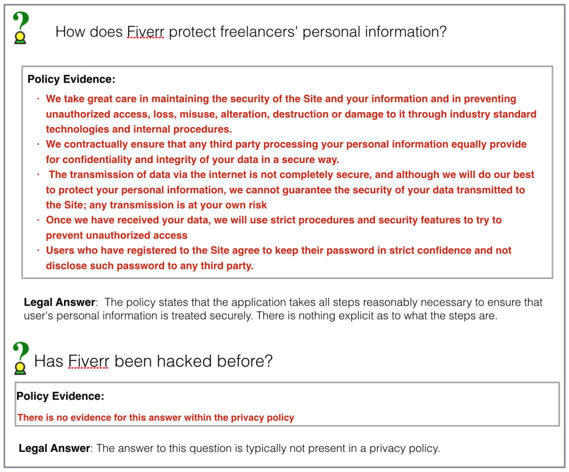
Abhilasha Ravichander[§], **Alan Black**[§], **Eduard Hovy**[§],
Joel Reidenberg[†], **N. Cameron Russell**[†], **Norman Sadeh**[§]

[§] Carnegie Mellon University
School of Computer Science
Pittsburgh, PA, USA
{aravicha, awb, ehovy, sadeh}@cs.cmu.edu

[†] Fordham University
Law School,
New York, NY, USA
jreidenberg@law.fordham.edu

Abstract

Privacy policies are legal documents used to inform users about the collection and handling of their data services or technologies with which they interact. Research has shown that few users take the time to read these policies, as they are often long and difficult to understand. In addition, users often only care about a small subset of issues discussed in privacy policies, and some of the issues they actually care about may not even be addressed in the text of the policies. Rather than requiring users to read the policies, a better approach might be to allow them to simply ask questions about those issues they care about, possibly through iterative dialog. In this work, we take a step towards this goal by exploring the idea of an automated privacy question-answering assistant, and look at the kinds of questions users are likely to pose to such a system. This analysis is informed by an initial study that elicits privacy questions from crowdworkers about the data practices of mobile apps. We analyze 1350 questions posed by crowdworkers about the privacy practices of a diverse cross section of mobile applications. This analysis sheds some light on privacy issues mobile app users are likely to inquire about as well as their ability to articulate questions in this domain. Our findings in turn should help inform the design of future privacy question answering systems.



How does Fiverr protect freelancers' personal information?

Policy Evidence:

- We take great care in maintaining the security of the Site and your information and in preventing unauthorized access, loss, misuse, alteration, destruction or damage to it through industry standard technologies and internal procedures.
- We contractually ensure that any third party processing your personal information equally provide for confidentiality and integrity of your data in a secure way.
- The transmission of data via the internet is not completely secure, and although we will do our best to protect your personal information, we cannot guarantee the security of your data transmitted to the Site; any transmission is at your own risk
- Once we have received your data, we will use strict procedures and security features to try to prevent unauthorized access
- Users who have registered to the Site agree to keep their password in strict confidence and not disclose such password to any third party.

Legal Answer: The policy states that the application takes all steps reasonably necessary to ensure that user's personal information is treated securely. There is nothing explicit as to what the steps are.

Has Fiverr been hacked before?

Policy Evidence:

There is no evidence for this answer within the privacy policy

Legal Answer: The answer to this question is typically not present in a privacy policy.

Figure 1: Examples of privacy-related questions users ask for *Fiverr*. Policy evidence represents sentences in the privacy policy that are relevant for determining the answer to the user's question.

Introduction

Privacy policies are the legal documents which disclose the ways in which a company gathers, uses, shares and manages user data. They are now nearly ubiquitous on websites and mobile applications. Privacy policies work under the ”notice and choice” regime, where users read privacy policies and can then choose whether or not to accept the terms of the policy, occasionally subject to some opt-in or opt-out provisions.

However due to the length and verbosity of these documents (Cate, 2010; Cranor, 2012; Schaub et al., 2015; Gluck et al., 2016), the average user does not read the privacy policies they consent to (Jain, Gyanchandani, and Khare, 2016; Commission and others, 2012). McDonald and Cranor (2008) find that if users spent time reading the privacy policies for all the website they interact with, it would account for a significant portion of the time they currently spend on the web. This disconnect between the requirements of real Internet users and their theoretical behavior under the notice and choice paradigm render this model largely ineffective (Reidenberg et al., 2015b). This is an opportunity for language technologies to help better serve the needs of users, by processing privacy policies automatically and allowing users to engage with them through interactive dialog. The legal domain has long served as a useful application domain for Natural Language Processing techniques (Mahler, 2015), however the sheer pervasiveness of websites and mobile applications in today's world necessitates the creation of automatic techniques to help users better understand the content of privacy policies.

In this work, we explore the idea of an automated "privacy assistant", which allows users to explore the content of a privacy policy by answering their questions. This kind of question-answering approach would allow for a more personalized approach to privacy, enabling users to review sections of policies that they are most interested in. The successful development of effective question-answering functionality for privacy requires a careful understanding of the types of questions users are likely to ask, how users are likely to formulate these questions, as well as estimating the difficulty of answering these questions. In this work, it is our goal to explore these issues by providing a preliminary qualitative analysis of privacy-related questions posed by crowdworkers.

Related Work

Policy Analysis

There has been considerable interest in making the content of privacy policies easy to understand. These include approaches that prescribe guidelines for drafting privacy policies (Kelley et al., 2009; Micheti, Burkell, and Steeves, 2010) or require service providers to encode privacy-policies in a machine-readable format (Cranor, 2003). These methods have not seen widespread adoption from industry and were abandoned. More recently, the community has been looking at automatically understanding the content of privacy policies (Sadeh et al., 2013; Liu et al., 2016; Oltramari et al., 2017; Mysore Sathyendra et al., 2017; Wilson et al., 2017). Perhaps most closely related to our contribution is the work of Harkous et al. (2018), which investigates answering questions from privacy policies by looking at privacy-related questions users ask companies on Twitter and annotating "segments" in the privacy policy as being relevant answers. Our study differs from their approach in several ways. First, our study is an order larger in magnitude . This is in part due to the scalability of our crowdsourcing methodology (§), at the expense of having 'natural' questions. However, as we show later in this work finding such questions in the wild can also be challenging. Secondly, we take into account for the fact that an answer to a question might not always be in the privacy policy, and if it is, it is possible there are multiple correct answers. This more accurately reflects a real-world scenario where users can ask any question of a privacy assistant. Third, our answers are provided by domain experts with legal training. Moreover, the annotations are provided at a sentence-level granularity. This is a considerable advantage over segment-level annotations for two reasons: first, the concept of what constitutes a segment is poorly defined and has different meanings to different audiences whereas the notion of a sentence is much more objective. Second : a finer level of granularity allows us to eliminate redundant information within segments, and presenting irrelevant information to a user detracts from how helpful an answer is. A system can always default to presenting segment-level information if required, by selecting all the sentences within the segment. Sathyendra et al. (2017) present some initial approaches to question answering for privacy policies. They outline several avenues for future work, including the need to elicit more representative datasets, determine if questions are unanswerable, and decrease reliance on segments. Our work takes a first step in this direction through a crowdsourced study that elicits a wide range of questions as well as legally-sound answers at the sentence-level of granularity.

Reading Comprehension

Several large-scale reading comprehension/answer selection datasets exist for Wikipedia passages (Rajpurkar et al., 2016; Rajpurkar, Jia, and Liang, 2018; Joshi et al., 2017; Choi et al., 2018) and news articles (Trischler et al., 2016; Hermann et al., 2015; Onishi et al., 2016). Our work considers question-answering within the specialized privacy domain, where documents are typically long and complex, and their accurate interpretation requires legal expertise. Thus, our work can also be considered to be related to similar efforts in the legal domain (e.g., Monroy, Calvo, and Gelbukh (2009); Quaresma and Rodrigues (2005)). These approaches are based on information retrieval for legal documents and have primarily been applied to juridical documents. Do et al. (2017) describes retrieving relevant Japanese Civil Code documents for question answering. Kim, Xu, and Goebel (2015) investigate answering true/false questions from Japanese bar exams. Liu, Chen, and Ho (2015) explores finding relevant Taiwanese legal statutes for a natural language query . A number of authors have also described domain-specific knowledge engineering approaches combining ontologies and knowledge bases to answer questions (e.g., Mollá and Vicedo (2007); Frank et al. (2007)). Feng et al. (2015); Tan et al. (2016) look at non-factoid question answering in the insurance domain. Each of these specialized domains present their own unique challenges, and progress in them requires a careful understanding of the domain as well as best practices in presenting information to the end user.

Crowdsourced Study

We would like to gain a better understanding of the kinds of questions users are likely to ask, and what legally-sound answers to them would be. For this purpose, we collect our data in two stages: first, we crowdsource questions on the contents of privacy policies from crowdworkers, and then we rely on domain experts with legal training to provide answers to the questions. We would like to note that our methodology only exposes crowdworkers to public information about each of the companies, rather than requiring them to read the privacy policy to formulate questions. This includes the name of the mobile application, the description of the mobile application as presented on the Google Playstore as well as screenshots from the mobile application. This approach attempts to circumvent potential bias from lexical entrainment, and more generally the risk of biasing crowdworkers to ask questions only about the practices disclosed in the privacy policy.

In this study we intentionally select mobile applications from a number of different categories, specifically focusing on apps from categories that occupy $\geq 2\%$ of mobile applications on the Google Playstore (Story, Zimmeck, and Sadeh, 2018)[1] [2] [3]. We would like to collect a representative

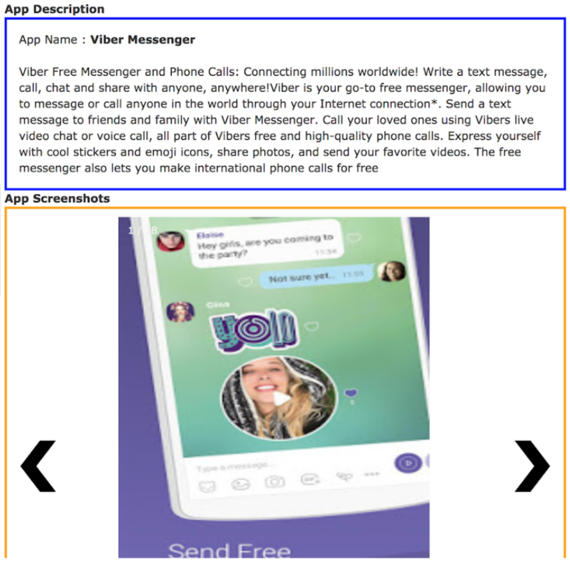
[1] As of April 1, 2018

[2] Games are by far the largest category of apps on the Google Playstore. We collapse the different game subcategories into one category for our purposes.

[3] We choose to focus on the privacy policies of mobile applications given the ubiquitousness of smartphones. However, our study design is limited to Android mobile applications. In practice however, these mobile applications often share privacy policies across platforms

Statistic	Train	Test	All
# Questions	1000	350	1350
# Passages	20	7	27
# Sentences	2879	909	3788
Avg Question Length	8.44	8.56	8.47
Avg Passage Length	3372.1	2990.29	3273.11
Avg Answer Length	93.94	111.9	104.52

Table 1: Statistics of PrivacyQA Dataset, where # denotes number of questions, passages and sentences, and average length of questions, passages and answers in words, for training and test partitions.



App Description

App Name : **Viber Messenger**

Viber Free Messenger and Phone Calls: Connecting millions worldwide! Write a text message, call, chat and share with anyone, anywhere!Viber is your go-to free messenger, allowing you to message or call anyone in the world through your Internet connection*. Send a text message to friends and family with Viber Messenger. Call your loved ones using Vibers live video chat or voice call, all part of Vibers free and high-quality phone calls. Express yourself with cool stickers and emoji icons, share photos, and send your favorite videos. The free messenger also lets you make international phone calls for free

App Screenshots

Figure 2: User interface for question elicitation.

set of questions such that we range from mobile applications which are well-known and likely to have carefully constructed privacy policies, all the way to applications which may have smaller install bases and less sophisticated privacy policies. We sample applications from each category using the Google Playstore recommendation engine, such that only half of the applications in our corpus have more than 5 million installs [4]. We collect data for 27 privacy policies across 10 categories of mobile applications. [5]

Crowdsourced Question Elicitation

An important objective of this study is to elicit and understand the types of questions users are likely to have when looking to install a mobile application. As discussed earlier, we present information similar to the information found when looking at the application in the Google playstore (Figure 2). We use Amazon Mechanical Turk to elicit questions about these privacy policies. Crowdworkers were asked to imagine they installed a mobile application and could talk to a trusted privacy assistant, whom they could ask any privacy-related question pertaining to the app. They were paid 12$ per hour to ask five questions for a given policy. We solicited questions from Turkers who were conferred "master" status, and whose location was within the United States and our task received favorable reviews on TurkerHub. For each mobile application, crowdworkers were also asked to rate their understanding of what the app does on a Likert scale of 1-5 (ranging from not being familiar to understanding it extremely well), as well as to indicate whether they had installed or used the app before. We also collected demographic information regarding the age of the crowdworkers.

Answer Selection

We are not just interested in collecting data on what questions users ask, but also a corpus of what good answers to these questions would be. For this purpose, given questions for a particular application, we recruit four experts with legal training to formulate answers to these questions based on the text of that application's privacy policy. The experts annotate questions for their relevance, subjectivity and also identify the relevant OPP-115(Wilson et al., 2016) category(ies) corresponding to each question, if any. We then formulate the problem of answering the question as a sentence selection task, and ask our annotators to find supporting evidence in the document which can help in answering the question. In this way, every question is shown to at least one annotator, and 350 questions are annotated by multiple annotators [6].

Analysis

Table 1. describes the results of our data collection effort. We receive 1350 questions to our imaginary privacy assistant, about the privacy practices of 27 mobile applications. The question length is on average 8.4 words and the privacy policies are typically very long pieces of text, 3̃000 words long. The answers to the questions typically have 1̃00 words of evidence in the privacy policy document.

What types of questions do users ask the privacy assistant?

We would like to explore the kinds of questions that users ask our conversational assistant. We analyze questions based on their question words, as well as by having our expert annotators indicate whether they believe the questions are related to privacy, whether they are subjective in nature and what categories they belong to in the OPP-15 ontology (Wilson et al., 2016). The results of this analysis is as follows[7]:

[4] We choose 5 million installs as a threshold on popularity of the mobile application, but this choice is debatable. Mobile applications with fewer than 5 million installs could also represent applications of large corporations and vice versa.

[5] The Playstore categories we sample applications from include: Books and Reference, Business, Education, Entertainment, Lifestyle, Health and Fitness, News and Magazines, Tools, Travel and Local, and Games.

[6] These form our held-out test set.

[7] All analyses in this section are presented on the 'All' data split unless mentioned otherwise

Question Word	Percentage
is/does	27.9 %
what	13.5 %
will	11.9 %
how	10.1 %
can	8.6 %
are	4.5 %
who	4.4 %
where	1.3 %
if	1.8 %

Table 2: Analysis of questions by question words for categories that account for >1% of questions

Question Words We qualitatively analyze questions by their expected types, based on the first word of the question. Note that while the question word can give us some information about the information-seeking intent of the user, the different question words can often be used interchangeably. For example, the questions *'will it collect my location?'* can also be phrased as *'does it collect my location'*. Keeping these limitations in mind, we perform a qualitative analysis of the elicited questions to identify common user intents. The distributions of questions across types can be found in Table 2. By far, the largest proportion of questions can be grouped into the 'is/does' category where, similar to the 'are' category, users are often questioning the assistant about a particular privacy attribute (for example, *'does this app track my location?'* or *'is this app tracking my location?'*). The next largest category includes 'what' questions which include a broad spectrum of questions (for example, *what sort of analytics are integrated in the app?'* or *'what do you do with my information'*). The 'will' and 'can' questions are usually asking about a potential privacy harm (for example, *'will i be explicitly told when my info is being shared with a third party?'* or *'will any academic institutions or employers be able to access my performance/score information?'* or *'can the app see what i type and what i search for?'*). 'How' questions generally either ask about specific company processes, or abstract attributes, such as security, longevity of data retention etc (for example, *'how safe is my password'* and *'how is my data protected'*). Relevant 'where' questions are generally related to data storage (for example, *'Where is my data stored?'*). Questions that begin with 'who' are usually asking about first party or third party access to data (for example, *'who can see my account information?'* or *'who all has access to my medical information?'*). Finally questions in the 'if' category typically establish a premise, before asking a question. Such a question needs to be answered based on both the contents of the policy as well as assuming the information in the premise is true (for example, *'if i link it to my Facebook will it have access to view my private information?'* or *'if i choose to opt out of the app gathering my personal data, can i still use the app?'*).

Relevance and Subjectivity We analyze how many of the questions asked to our privacy assistant are 'relevant' i.e are related to privacy, and how many are subjective in nature. In the real-world it isn't necessary that users will only ask our privacy assistant questions related to privacy. Thus, it is important for us to be able to identify which questions we are capable of attempting to answer. We analyze the test-set where each example features multiple annotations from our expert-annotators. We consider the majority-vote to be the judgement of whether a question is relevant or subjective. We find that 78.85% of questions received by our privacy assistant are relevant, with 6.28% being subjective. Table. 3 gives us more insight into this phenomena. We observe that the majority of questions (74%) are relevant but not subjective (for example, *'what information are they collecting?'*). 4.86% of questions are both relevant and subjective (for example, *'is my data safe?'*), 1.4% are subjective but not relevant (for example, *'are there any in game purchases in the wordscapes app that i should be concerned about?'*) and finally 19.71% are neither relevant nor subjective (*'does the app require an account to play?'*).

Property	Privacy-Related	Not Privacy-Related
Subjective	4.86%	1.43%
Not Subjective	74%	19.71%

Table 3: Relevance and subjectivity judgments for 350 questions posed by crowdworkers.

Question Ontology Categories Next we ask our annotators to indicate the OPP-15 data practice category (Wilson et al., 2016) that best describes the question. Broadly, the ontology describes 10 data practice categories. The interested reader is invited to refer to (Wilson et al., 2016) for a detailed description of these data practices. Annotators are allowed to annotate a question as belonging to multiple categories. For example, the question *'What information of mine is collected by this app and who is it shared with?'* might belong to both the 'First Party Collection and Use' and the 'Third Party Sharing and Collection' OPP-115 data practice categories. We consider a category to be correct, if at least 2 annotators identify it to be relevant. In cases where none of the categories are identified as relevant, we default to 'other' if it is identified as a relevant category by at least one annotator. If not, we mark the category as 'no agreement'. The results from this analysis are presented in Table. 4. We observe that questions about first party and third party practices account for nearly 58.7% of all the questions asked of our assistant.

Comparative Analysis We analyze 100 samples drawn from the Twitter privacy dataset (Harkous et al., 2018), annotating them for OPP-category, relevance and if they are a question or not. We find that in the Twitter dataset, 23 % of the questions are complaints rather than questions. By OPP-category classification, 26% are First Party, 37% are Third Party, 14% are Data Security, 5% are User Access, 3% are User choice and 9% could be grouped in the 'other' category. Only 6% of the questions collected are not privacy related.

Privacy Practice	Percentage	Example
First Party Collection//Use	36.4 %	what data does this game collect?
Third Party Sharing//Collection	22.3 %	will my data be sold to advertisers?
Data Security	10.9 %	how is my info protected from hackers?
Data Retention	4.2 %	how long do you save my information?
User Access, Edit and Deletion	2.6 %	can i delete my information permanently?
User Choice//Control	7.2 %	is there a way to opt out of data sharing
Other	9.4 %	does the app connect to the internet at any point during its use?
International and Specific Audiences	0.6 %	what are your GDPR policies?
No Agreement	6.6 %	how are features personalized?

Table 4: OPP-115 categories most relevant to the questions collected from users.

Experiments

We would like to characterize and study the difficulty of the question-answering task for humans. We formulate the problem of identifying relevant evidence in the document to answer the question as a sentence-selection task, where it is possible to choose not to answer a question by not identifying any relevant sentences. We evaluate using sentence-level F1 rather than IR-style metrics so as to accommodate models to abstain from answering [8]. Similar to Choi et al. (2018); Rajpurkar et al. (2016), we compute the maximum F1 amongst all the reference answers. As abstaining from giving an answer is always legally sound but seldom helpful, we do not consider a question to be unanswerable if only a minority of experts abstain from giving an answer. Similar to (Choi et al., 2018) given n reference answers, we report the average maximum F1 performance of the $(n-1)$th subset compared to the heldout reference.

As discussed previously, since most questions are difficult to answer in a legally-sound way based on the contents of the privacy policy alone, abstaining from answering is often going to be a safe action. We would like to emphasize that this is not a criticism of the annotators or the people asking the questions, but rather a characteristic of this domain where privacy policies are often silent or ambiguous on issues users are likely to inquire about. To quantify the magnitude of this effect, we demonstrate that a model which always abstains from answering the question can achieve reasonable performance (Table 5), yet still leaves a large gap for improvement. We would further like to understand what makes the majority of our annotators decide a question should not be answered. We randomly sample 100 questions that were deemed unanswerable, and annotate them post-hoc with reasons informed by expert annotations. We find that for 56% of unanswerable questions, the answer to the question would typically not be present in most privacy policies. These would include questions such as '*how does the currency within the game work?*' and suggests that users would benefit from being informed about the scope of typical privacy policies. However, they also include questions such as '*has Viber had data breaches in the past?*' which ideally a privacy assistant would be able to answer, but is not present within a typical privacy policy. In the future, a privacy assistant could draw upon various sources of information such as metadata from the Google Playstore, background legal knowledge, news articles, social media etc. in order to broaden its coverage across questions. For an additional 24% of unanswerable questions, the answers were expected to be found in the privacy policy, but the privacy policy was silent on a possible answer (such as '*is my app data encrypted?*'). Generally when a policy is silent it is not safe to make any assumptions. 6% of questions asked by a user are too vague to understand correctly such as *'who can contact me through the app?'*, such questions would benefit from the assistant engaging in a clarification dialogue. Another 4% are ambiguously phrased, such as *'any difficulties to occupy the privacy assistant?'*. These kind of questions are very hard to interpret correctly. 3% of unanswerable questions are *too specific* in nature, and it is unlikely the creators of the privacy policy would anticipate that particular question ('*does it have access to financial apps i use?*'). Finally, 7% of unanswerable questions are too subjective and our annotators tend to abstain from answering (for example, *'how do i know this app is legit?'*).

Model	Precision	Recall	F1
No Answer (NA)	36.2	36.2	36.2
Human	70.3	71.1	70.7

Table 5: Human performance and performance of a No-Answer Baseline. Human performance demonstrates considerable agreement on the right answer for the privacy domain, where experts often disagree (Reidenberg et al., 2015a).

We would also like to be able to characterize the disagreement on this task. It is important to note here that all of our annotators are experts with legal training rather than crowdworkers, and their provided answers can generally be assumed to be valid legal opinions about the question. We tease apart the difference from where they abstained to answer to their disagreements by comparing against the No Answer (henceforth known as NA) baseline (Table 5). In Table 5 we observe the human F1 is 70.7%, demonstrating considerable agreement on the right answer. We would still like

[8] Similar to (Rajpurkar, Jia, and Liang, 2018) and (Yang, Yih, and Meek, 2015), for negative examples models are awarded 1 F1 if they abstain from answering and 0 F1 for any answer at all

Question Word	NA Model	Human
is/does	37.22	73.19
what	39.77	73.35
will	13.04	66.56
how	27.84	80.16
can	27.17	63.04
are	35.85	68.68
who	17.02	58.44
where	54.55	54.55
if	0	62.19

Table 6: Classifier performance in F1 stratified by first word in the question.

Privacy Practice	NA Model	Human
First Party Collection/Use	24.6	67.1
Third Party Sharing/Collection	6.9	60.6
Data Security	35.3	87.2
Data Retention	0	79.8
User Access, Edit and Deletion	0	53.1
User Choice/Control	46.3	64.7
Other	89.1	84.1
International & Specific Audiences	0	100
No Agreement	76.2	78.3

Table 7: Classifier performance in F1 stratified by OPP-115 category of the question.

to investigate whether any disagreements are valid, or if they are due to poor definitions or lack of adequate specification in the annotation instructions. We randomly sample 50 samples and annotate them for likely reasons for disagreement [9]. We find that they "agree on 64% of instances and disagree on 36%. We further determine that 92.8% of disagreements were legitimate, valid different interpretations. For 43.75% the question was interpreted differently, in 25% the contents of the privacy policy were interpreted differently and the remaining were due to other sources of error (for example, in the question *'who is allowed to use the app'*, most annotators abstain from answering, but one annotator points out that the policy states that children under the age of 13 are *not* allowed to use the app.)

We next analyze disagreements based on the type of question that was asked (Table. 6). As observed, the wh-type of the question may give us some information about the intent of the questions. We observe that our expert annotators rarely abstain to answer when a user asks a 'will' question about a potential privacy harm, taking care to identify relevant sections of the privacy policy. Similarly 'if' type questions generally are quite specific and require careful reasoning. On the other hand 'where' questions are generally about data storage. They are vague, for example *'where is my data stored?'* is probably not asking for the exact location of the company's datacenters but it is unclear what granularity is meant in the question (e.g., particular country, versus knowing whether the data is stored on a mobile phone or in the cloud).

We also analyze disagreements based on the OPP-115 category of the question (Table. 7). As expected, questions where annotators disagree on the category of the question, have more disagreements than simply abstaining to answer. Similarly for user choice, the policy typically does not answer questions like *'how do I limit its access to data'* fully, so the annotators tend to abstain from answering. In contrast, questions about first party and third party practices are usually anticipated and often have answers in the privacy policy.

Conclusion

What kinds of questions should an automated privacy assistant expect to receive? We explore this question by designing a study that elicits questions from crowdworkers who are asked to think about the data practices of mobile apps they might consider downloading on their smartphones. We qualitatively analyze the types of questions asked by users, and identify a number of challenges associated with generating answers to these questions. While in principle privacy policies should be written to answer questions users are likely to have, in practice, our study shows that questions asked by users often go beyond what is disclosed in the text of privacy policies. Challenges arise in automated question answering, both because policies are often silent or ambiguous on issues that users are likely to inquire about, and also because users are not very good at articulating their privacy questions - and occasionally even ask questions that have nothing to do with privacy. Determining a user's intent may be a process of discovery for both the user and the assistant, and thus in the future it would be helpful if the assistant was capable of engaging in clarification dialogue. Such a privacy assistant would have to reconcile the need to be helpful to the user and provide answers that are legally accurate with the need to be helpful. It would have to be capable of disambiguating questions by engaging in dialogues with users; it would have to be able to supplement information found (or lacking) in the privacy policy with additional sources of information such as background legal knowledge. Ideally, it would also be able to interpret ambiguity in the policy and also be able interpret silence about different issues. We hope that the identification of these requirements will help inform the design of effective automatic privacy assistants.

Acknowledgements

This work has been supported by the National Science Foundation under Grant No. CNS 13-30596 and No. CNS CNS 13-30214. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, or the US Government. The authors would like to thank Lorrie Cranor, Florian Schaub and Shomir Wilson for insightful feedback and discussion related to this work.

[9] We do not use F1 to measure disagreement, and instead manually filter samples so we can capture both when the legal experts interpreted the question differently, as well as when they interpret the contents of the privacy policy differently.

References

Cate, F. H. 2010. The limits of notice and choice. *IEEE Security & Privacy* 8(2):59–62.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Commission, U. F. T., et al. 2012. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. *FTC Report*.

Cranor, L. F. 2003. P3p: Making privacy policies more useful. *IEEE Security & Privacy* 99(6):50–55.

Cranor, L. F. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10:273.

Do, P.-K.; Nguyen, H.-T.; Tran, C.-X.; Nguyen, M.-T.; and Nguyen, M.-L. 2017. Legal question answering using ranking svm and deep convolutional neural network. *arXiv preprint arXiv:1703.05320*.

Feng, M.; Xiang, B.; Glass, M. R.; Wang, L.; and Zhou, B. 2015. Applying deep learning to answer selection: A study and an open task. *arXiv preprint arXiv:1508.01585*.

Frank, A.; Krieger, H.-U.; Xu, F.; Uszkoreit, H.; Crysmann, B.; Jörg, B.; and Schäfer, U. 2007. Question answering from structured knowledge sources. *Journal of Applied Logic* 5(1):20–48.

Gluck, J.; Schaub, F.; Friedman, A.; Habib, H.; Sadeh, N.; Cranor, L. F.; and Agarwal, Y. 2016. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *12th Symposium on Usable Privacy and Security (SOUPS)*, 321–340.

Harkous, H.; Fawaz, K.; Lebret, R.; Schaub, F.; Shin, K. G.; and Aberer, K. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. *arXiv preprint arXiv:1802.02561*.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, 1693–1701.

Jain, P.; Gyanchandani, M.; and Khare, N. 2016. Big data privacy: a technological perspective and review. *Journal of Big Data* 3(1):25.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kelley, P. G.; Bresee, J.; Cranor, L. F.; and Reeder, R. W. 2009. A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, 4. ACM.

Kim, M.-Y.; Xu, Y.; and Goebel, R. 2015. Applying a convolutional neural network to legal question answering. In *JSAI International Symposium on Artificial Intelligence*, 282–294. Springer.

Liu, F.; Wilson, S.; Schaub, F.; and Sadeh, N. 2016. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *2016 AAAI Fall Symposium Series*.

Liu, Y.-H.; Chen, Y.-L.; and Ho, W.-L. 2015. Predicting associated statutes for legal problems. *Information Processing & Management* 51(1):194–211.

Mahler, L. 2015. What is nlp and why should lawyers care. *Retrieved March* 12:2018.

McDonald, A. M., and Cranor, L. F. 2008. The cost of reading privacy policies. *ISJLP* 4:543.

Micheti, A.; Burkell, J.; and Steeves, V. 2010. Fixing broken doors: Strategies for drafting privacy policies young people can understand. *Bulletin of Science, Technology & Society* 30(2):130–143.

Mollá, D., and Vicedo, J. L. 2007. Question answering in restricted domains: An overview. *Computational Linguistics* 33(1):41–61.

Monroy, A.; Calvo, H.; and Gelbukh, A. 2009. Nlp for shallow question answering of legal documents using graphs. *Computational Linguistics and Intelligent Text Processing* 498–508.

Mysore Sathyendra, K.; Wilson, S.; Schaub, F.; Zimmeck, S.; and Sadeh, N. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2774–2779. Copenhagen, Denmark: Association for Computational Linguistics.

Oltramari, A.; Piraviperumal, D.; Schaub, F.; Wilson, S.; Cherivirala, S.; Norton, T. B.; Russell, N. C.; Story, P.; Reidenberg, J.; and Sadeh, N. 2017. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web* (Preprint):1–19.

Onishi, T.; Wang, H.; Bansal, M.; Gimpel, K.; and McAllester, D. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2230–2235. Austin, Texas: Association for Computational Linguistics.

Quaresma, P., and Rodrigues, I. P. 2005. A question answer system for legal information retrieval. In *JURIX*, 91–100.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Reidenberg, J. R.; Breaux, T.; Cranor, L. F.; French, B.; Grannis, A.; Graves, J. T.; Liu, F.; McDonald, A.; Norton, T. B.; and Ramanath, R. 2015a. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ* 30:39.

Reidenberg, J. R.; Russell, N. C.; Callen, A. J.; Qasir, S.; and Norton, T. B. 2015b. Privacy harms and the effectiveness of the notice and choice framework. *ISJLP* 11:485.

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.; et al. 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, Technical Report, CMU-ISR-13-119, Carnegie Mellon University.

Sathyendra, K. M.; Ravichander, A.; Story, P. G.; Black, A. W.; and Sadeh, N. 2017. Helping users understand privacy notices with automated query answering functionality: An exploratory study. *Technical Report*.

Schaub, F.; Balebako, R.; Durity, A. L.; and Cranor, L. F. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, 1–17.

Story, P.; Zimmeck, S.; and Sadeh, N. 2018. Which apps have privacy policies?

Tan, M.; dos Santos, C.; Xiang, B.; and Zhou, B. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 464–473. Berlin, Germany: Association for Computational Linguistics.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Wilson, S.; Schaub, F.; Dara, A. A.; Liu, F.; Cherivirala, S.; Leon, P. G.; Andersen, M. S.; Zimmeck, S.; Sathyendra, K. M.; Russell, N. C.; et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 1330–1340.

Wilson, S.; Schaub, F.; Liu, F.; Sathyendra, K.; Zimmeck, S.; Ramanath, R.; Liu, F.; Sadeh, N.; and Smith, N. 2017. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web*.

Yang, Y.; Yih, W.-t.; and Meek, C. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2013–2018.