Question Answering for Privacy Policies: Combining Computational and Legal Perspectives

Abhilasha Ravichander ♢ **Alan Black** ♢ **Shomir Wilson** ♡
Thomas Norton♠ **Norman Sadeh**♢
♢Carnegie Mellon University, Pittsburgh, PA
♡Penn State University, University Park, PA ♠Fordham Law School, New York, NY
{aravicha, awb, sadeh}@cs.cmu.edu
{shomir}@psu.edu, {tnorton1}@law.fordham.edu

Abstract

Privacy policies are long and complex documents that are difficult for users to read and understand, and yet, they have legal effects on how user data is collected, managed and used. Ideally, we would like to empower users to inform themselves about issues that matter to them, and enable them to selective explore those issues. We present PRIVACYQA, a corpus consisting of 1750 questions about the privacy policies of mobile applications, and over 3500 expert annotations of relevant answers. We observe that a strong neural baseline underperforms human performance by almost 0.3 F1 on PRIVACYQA, suggesting considerable room for improvement for future systems. Further, we use this dataset to shed light on challenges to *question answerability*, with domain-general implications for any question answering system. The PRIVACYQA corpus offers a challenging corpus for question answering, with genuine real-world utility.

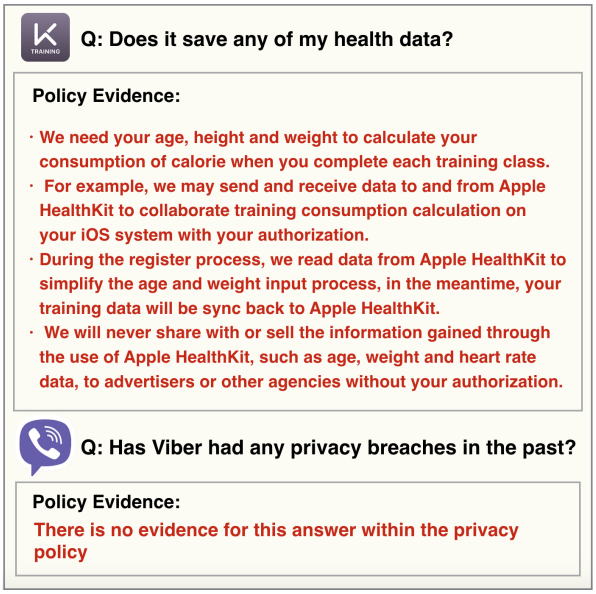

Figure 1: Examples of privacy-related questions users ask, drawn from two mobile applications: Keep[1] and Viber.[2] Policy evidence represents sentences in the privacy policy that are relevant for determining the answer to the user's question.[3]

1 Introduction

Privacy policies are the documents which disclose the ways in which a company gathers, uses, shares and manages a user's data. As legal documents, they function using the principle of *notice and choice* (Federal Trade Commission, 1998), where companies post their policies, and theoretically, users read the policies and decide to use a company's products or services only if they find the conditions outlined in its privacy policy acceptable. Many legal jurisdictions around the world accept this framework, including the United States and the European Union (Patrick, 1980; OECD, 2004). However, the legitimacy of this framework depends upon users actually reading and understanding privacy policies to determine whether company practices are acceptable to them (Reidenberg et al., 2015). In practice this is seldom the case (Cate, 2010; Cranor, 2012; Schaub et al., 2015; Gluck et al., 2016; Jain et al., 2016; US Federal Trade Commission et al., 2012; McDonald and Cranor, 2008). This is further complicated by the highly individual and nuanced compromises that users are willing to make with their data (Leon et al., 2015), discouraging a 'one-size-fits-all' approach to notice of data practices in privacy documents.

With devices constantly monitoring our environment, including our personal space and our bodies, lack of awareness of how our data is being used easily leads to problematic situations where users are outraged by information misuse, but companies insist that users have consented. The discovery of increasingly egregious uses of data by companies, such as the scandals involv-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4947–4958, Hong Kong, China, November 3–7, 2019.

ing Facebook and Cambridge Analytica (Cadwalladr and Graham-Harrison, 2018), have further brought public attention to the privacy concerns of the internet and ubiquitous computing. This makes privacy a well-motivated application domain for NLP researchers, where advances in enabling users to quickly identify the privacy issues most salient to them can potentially have large real-world impact.

Motivated by this need, we contribute PRIVACYQA, a corpus consisting of 1750 questions about the contents of privacy policies[4], paired with over 3500 expert annotations. The goal of this effort is to kickstart the development of question-answering methods for this domain, to address the (unrealistic) expectation that a large population should be reading many policies per day. In doing so, we identify several understudied challenges to our ability to answer these questions, with broad implications for systems seeking to serve users' information-seeking intent. By releasing this resource, we hope to provide an impetus to develop systems capable of language understanding in this increasingly important domain.[5]

2 Related Work

Prior work has aimed to make privacy policies easier to understand. Prescriptive approaches towards communicating privacy information (Kelley et al., 2009; Micheti et al., 2010; Cranor, 2003) have not been widely adopted by industry. Recently, there have been significant research effort devoted to understanding privacy policies by leveraging NLP techniques (**?**Liu et al., 2016; Oltramari et al., 2017; Mysore Sathyendra et al., 2017; Wilson et al., 2017), especially by identifying specific data practices within a privacy policy. We adopt a personalized approach to understanding privacy policies, that allows users to query a document and selectively explore content salient to them. Most similar is the PolisisQA corpus (Harkous et al., 2018), which examines questions users ask corporations on Twitter. Our approach differs in several ways: 1) The PRIVACYQA dataset is larger, containing 10x as many questions and answers. 2) Answers are formulated by domain experts with legal training. [6] 3) PRIVACYQA includes diverse question types, including unanswerable and subjective questions.

Our work is also related to reading comprehension in the open domain, which is frequently based upon Wikipedia passages (Rajpurkar et al., 2016, 2018; Joshi et al., 2017; Choi et al., 2018) and news articles (Trischler et al., 2017; Hermann et al., 2015; Onishi et al., 2016). Table.1 presents the desirable attributes our dataset shares with past approaches. This work is also tied into research in applying NLP approaches to legal documents (Monroy et al., 2009; Quaresma and Rodrigues, 2005; Do et al., 2017; Kim et al., 2015; Liu et al., 2015; Mollá and Vicedo, 2007; Frank et al., 2007). While privacy policies have legal implications, their intended audience consists of the general public rather than individuals with legal expertise. This arrangement is problematic because the entities that write privacy policies often have different goals than the audience. Feng et al. (2015); Tan et al. (2016) examine question answering in the insurance domain, another specialized domain similar to privacy, where the intended audience is the general public.

3 Data Collection

We describe the data collection methodology used to construct PRIVACYQA. With the goal of achieving broad coverage across application types, we collect privacy policies from 35 mobile applications representing a number of different categories in the Google Play Store.[7][8] One of our goals is to include both policies from well-known applications, which are likely to have carefully-constructed privacy policies, and lesser-known applications with smaller install bases, whose policies might be considerably less sophisticated. Thus, setting 5 million installs as a threshold, we ensure each category includes applications with installs on both sides

[1] https://play.google.com/store/apps/details?id=com.gotokeep.keep.intl

[2] https://play.google.com/store/apps/details?id=com.viber.voip

[3] A question might not have any supporting evidence for an answer within the privacy policy.

[4] All privacy policies in this corpus are in English.

[5] PRIVACYQA is freely available at https://github.com/AbhilashaRavichander/PrivacyQA_EMNLP.

[6] This choice was made as privacy policies are legal documents, and require careful expert understanding in order to be interpreted correctly.

[7] We choose categories that occupy atleast a 2% share of all application categories on the Play Store (Story et al., 2018)

[8] As of April 1, 2018

Dataset	Document Source	Expert Annotator	Simple Evaluation	Unanswerable Questions	Asker Cannot See Evidence
PrivacyQA	Privacy Policies	✓	✓	✓	✓
NarrativeQA (Kočiskỳ et al., 2018)	Fiction	✗	✗	✗	✓
InsuranceQA (Feng et al., 2015)	Insurance	✓	✓	✗	✓
TriviaQA (Joshi et al., 2017)	Wikipedia	✗	✓	✗	✓
SQuAD 1.0 (Rajpurkar et al., 2016)	Wikipedia	✗	✓	✗	✗
SQuAD 2.0 (Rajpurkar et al., 2018)	Wikipedia	✗	✓	✓	✗
MS Marco (Nguyen et al., 2016)	Web Documents	✗	✗	✓	✓
MC Test (Richardson et al., 2013)	Fiction	✗	✓	✗	✗
NewsQA (Trischler et al., 2017)	News Articles	✗	✓	✓	✓

Table 1: Comparison of the PRIVACYQA dataset to other question answering datasets. Expert annotator indicates domain expertise of the answer provider. Simple evaluation indicates the presence of an automatically calculable evaluation metric. Unanswerable questions indicates if the respective corpus includes unanswerable questions. 'Asker Cannot See Evidence' indicates that the asker of the question was not shown evidence from the document at the time of formulating questions.

Figure 2: User interface for question elicitation.

of this threshold.[9] All policies included in the corpus are in English, and were collected before April 1, 2018, predating many companies' GDPR-focused (Voigt and Von dem Bussche, 2017) updates. We leave it to future studies (Gallé et al., 2019) to look at the impact of the GDPR (e.g., to what extent GDPR requirements contribute to making it possible to provide users with more informative answers, and to what extent their disclosures continue to omit issues that matter to users).

3.1 Crowdsourced Question Elicitation

The intended audience for privacy policies consists of the general public. This informs the decision to elicit questions from crowdworkers on the contents of privacy policies. We choose not to show the contents of privacy policies to crowdworkers, a procedure motivated by a desire to avoid inadvertent biases (Weissenborn et al., 2017; Kaushik and Lipton, 2018; Poliak et al., 2018; Gururangan et al., 2018; Naik et al., 2018), and encourage crowdworkers to ask a variety of questions beyond only asking questions based on practices described in the document.

Instead, crowdworkers are presented with public information about a mobile application available on the Google Play Store including its name, description and navigable screenshots. Figure 2 shows an example of our user interface.[10] Crowdworkers are asked to imagine they have access to a trusted third-party privacy assistant, to whom they can ask any privacy question about a given mobile application. We use the Amazon Mechanical Turk platform[11] and recruit crowdworkers who have been conferred "master" status and are located within the United States of America. Turkers are asked to provide five questions per mobile application, and are paid $2 per assignment, taking ~eight minutes to complete the task.

[9] The final application categories represented in the corpus consist of books, business, education, entertainment, lifestyle, music, health, news, personalization, photography, productivity, tools, travel and game applications.

[10] Color blindness affects approximately 4.5% of the worlds population (Deeb, 2005) We design all our crowdworker user interfaces to accommodate red-green color blindness.

[11] https://www.mturk.com/

Word	(%)	Example Question From PRIVACYQA
Will	13.03	will my data be sold or available to be used by third party entities?
Do	2.70	do you sell any of my information
What	1.84	what is the worst case scenario of giving my information to this game?
Are	1.60	are there any safety concerns i should know about using this app?
Can/ could	1.47	can i control who sees my data?
How	1.40	how secure is my stored data?
Where	1.14	where is my account info and online activity stored?
Has	0.69	has there ever been a security breach?
If	0.61	if i delete the app will you keep my data?
Does	0.46	does the app save the addresses that i enter?

Table 2: Ten most frequent first words in questions in the PRIVACYQA dataset. We observe high lexical diversity in prefixes with 35 unique first word types and 131 unique combinations of first and second words.

3.2 Answer Selection

To identify legally sound answers, we recruit seven experts with legal training to construct answers to Turker questions. Experts identify relevant evidence within the privacy policy, as well as provide meta-annotation on the question's relevance, subjectivity, OPP-115 category (Wilson et al., 2016), and how likely any privacy policy is to contain the answer to the question asked.

3.3 Analysis

Table.4 presents aggregate statistics of the PRIVACYQA dataset. 1750 questions are posed to our imaginary privacy assistant over 35 mobile applications and their associated privacy documents. As an initial step, we formulate the problem of answering user questions as an extractive sentence selection task, ignoring for now background knowledge, statistical data and legal expertise that could otherwise be brought to bear. The dataset is partitioned into a training set featuring 27 mobile applications and 1350 questions, and a test set consisting of 400 questions over 8 policy documents. This ensures that documents in training and test splits are mutually exclusive. Every question is answered by at least one expert. In addition, in order to estimate annotation reliability and provide for better evaluation, every question in the test set is answered by at least two additional experts.

Table 2 describes the distribution over first words of questions posed by crowdworkers. We also observe low redundancy in the questions posed by crowdworkers over each policy, with each policy receiving ~49.94 unique questions despite crowdworkers independently posing questions. Questions are on average 8.4 words long. As declining to answer a question can be a legally sound response but is seldom practically useful, answers to questions where a minority of experts abstain to answer are filtered from the dataset. Privacy policies are ~3000 words long on average. The answers to the question asked by the users typically have ~100 words of evidence in the privacy policy document.

3.3.1 Categories of Questions

Questions are organized under nine categories from the OPP-115 Corpus annotation scheme (Wilson et al., 2016):

1. First Party Collection/Use: What, why and how information is collected by the service provider
2. Third Party Sharing/Collection: What, why and how information shared with or collected by third parties
3. Data Security: Protection measures for user information
4. Data Retention: How long user information will be stored
5. User Choice/Control: Control options available to users
6. User Access, Edit and Deletion: If/how users can access, edit or delete information
7. Policy Change: Informing users if policy information has been changed
8. International and Specific Audiences: Practices pertaining to a specific group of users
9. Other: General text, contact information or practices not covered by other categories.

For each question, domain experts indicate one or more[12] relevant OPP-115 categories. We mark a category as relevant to a question if it is identified as such by at least two annotators. If no such category exists, the category is marked

[12]For example, some questions such as *'What information of mine is collected by this app and who is it shared with?'* can be identified as falling under both first party data/collection and third party collection/sharing categories.

Privacy Practice	Proportion	Example Question From PRIVACYQA
First Party Collection/Use	41.9 %	what data does this game collect?
Third Party Sharing/Collection	24.5 %	will my data be sold to advertisers?
Data Security	10.5 %	how is my info protected from hackers?
Data Retention	4.1 %	how long do you save my information?
User Access, Edit and Deletion	2.0 %	can i delete my information permanently?
User Choice/Control	6.5 %	is there a way to opt out of data sharing
Other	4.8 %	does the app connect to the internet at any point?
Policy Change	0.2 %	where is the privacy statement
International and Specific Audiences	0.2 %	what are your GDPR policies?
No Agreement	5.4 %	how are features personalized?

Table 3: OPP-115 categories most relevant to the questions collected from users.

Dataset	Train	Test	All
# Questions	1350	400	1750
# Policies	27	8	35
# Sentences	3704	1243	4947
Avg Q. Length	8.42	8.56	8.46
Avg Doc. Length	3121.3	3629.13	3237.37
Avg Ans. Length	123.73	153.44	139.62

Table 4: Statistics of the PRIVACYQA Dataset, where # denotes number of questions, policies and sentences, and average length of questions, policies and answers in words, for training and test partitions.

as 'Other' if atleast one annotator has identified the 'Other' category to be relevant. If neither of these conditions is satisfied, we label the question as having no agreement. The distribution of questions in the corpus across OPP-115 categories is as shown in Table.3. First party and third party related questions are the largest categories, forming nearly 66.4% of all questions asked to the privacy assistant.

3.3.2 Answer Validation

When do experts disagree? We would like to analyze the reasons for potential disagreement on the annotation task, to ensure disagreements arise due to valid differences in opinion rather than lack of adequate specification in annotation guidelines. It is important to note that the annotators are experts rather than crowdworkers. Accordingly, their judgements can be considered valid, legally-informed opinions even when their perspectives differ. For the sake of this question we randomly sample 100 instances in the test data and analyze them for likely reasons for disagreements. We consider a disagreement to have occurred when more than one expert does not agree with the majority consensus. By disagreement we mean there is no overlap between the text identified as relevant by one expert and another.

We find that the annotators agree on the answer for 74% of the questions, even if the supporting evidence they identify is not identical i.e full overlap. They disagree on the remaining 26%. Sources of apparent disagreement correspond to situations when different experts: have differing interpretations of question intent (11%) (for example, when a user asks *'who can contact me through the app'*, the questions admits multiple interpretations, including seeking information about the features of the app, asking about first party collection/use of data or asking about third party collection/use of data), identify different sources of evidence for questions that ask if a practice is performed or not (4%), have differing interpretations of policy content (3%), identify a partial answer to a question in the privacy policy (2%) (for example, when the user asks *'who is allowed to use the app'* a majority of our annotators decline to answer, but the remaining annotators highlight partial evidence in the privacy policy which states that children under the age of 13 are *not* allowed to use the app), and other legitimate sources of disagreement (6%) which include personal subjective views of the annotators (for example, when the user asks 'is my DNA information used in any way *other than what is specified*', some experts consider the boilerplate text of the privacy policy which states that it abides to practices described in the policy document as sufficient evidence to answer this question, whereas others do not).

	Acc.	P	R	F1
Majority	24.75	24.75	100	39.6
SVM-BOW	75.75	50.8	58.5	54.4
+ LEN	76.75	52.7	57.5	55.0
+ LEN + POS	77.0	53.2	58.5	55.7
CNN	80.0	61.1	52.5	56.5
BERT	81.15	62.6	62.6	62.6

Table 5: Classifier Performance (%) for answerability of questions. The Majority Class baseline always predicts that questions are unanswerable.

Model	Precision	Recall	F1
No Answer (NA)	28.0%	28.0%	28.0%
Word Count -2	24.0%	16.4%	19.4%
Word Count -3	21.8%	17.8%	19.6%
Word Count -5	18.1%	20.3%	19.2%
BERT	44.2%	34.8%	39.0%
BERT + Unans.	44.3%	36.1%	**39.8**%
Human	68.8%	69.0%	68.9%

Table 6: Performance of baselines on PRIVACYQA dataset.

4 Experimental Setup

We evaluate the ability of machine learning methods to identify relevant evidence for questions in the privacy domain.[13] We establish baselines for the subtask of deciding on the answerability (§4.1) of a question, as well as the overall task of identifying evidence for questions from policies (§4.2). We describe aspects of the question that can render it unanswerable within the privacy domain (§5.2).

4.1 Answerability Identification Baselines

We define answerability identification as a binary classification task, evaluating model ability to predict if a question can be answered, given a question in isolation. This can serve as a prior for downstream question-answering. We describe three baselines on the answerability task, and find they considerably improve performance over a majority-class baseline.

SVM: We define 3 sets of features to characterize each question. The first is a simple bag-of-words set of features over the question (SVM-BOW), the second is bag-of-words features of the question as well as length of the question in words (SVM-BOW + LEN), and lastly we extract bag-of-words features, length of the question in words as well as part-of-speech tags for the question (SVM-BOW + LEN + POS). This results in vectors of 200, 201 and 228 dimensions respectively, which are provided to an SVM with a linear kernel.

CNN: We utilize a CNN neural encoder for answerability prediction. We use GloVe word embeddings (Pennington et al., 2014), and a filter size of 5 with 64 filters to encode questions.

BERT: BERT (Devlin et al., 2019) is a bidirectional transformer-based language-model (Vaswani et al., 2017).[14] We fine-tune BERT-base on our binary answerability identification task with a learning rate of 2e-5 for 3 epochs, with a maximum sequence length of 128.

Privacy Practice	NA	BERT-U	Human
First Party	0.22	0.36	0.67
Third Party	0.10	0.26	0.61
Data Security	0.24	0.42	0.74
Data Retention	0.02	0.33	0.67
User Access	0.07	0.32	0.66
User Choice	0.35	0.41	0.65
Other	0.44	0.45	0.72

Table 7: Stratification of classifier performance by OPP-115 category of questions.

	BERT
# Answerability Mistakes	137
% Answerable ->Unanswerable	124
% Unanswerable ->Answerable	13
Out-of-scope	2
Subjective	46
Policy Silent	19
Unexpected	6

Table 8: Analysis of BERT performance at identifying answerability. The majority of mistakes made by BERT are answerable Questions identified as unanswerable. These answerable questions are further analyzed along the factors of scope, subjectivity, presence of answer and whether the question could be anticipated.

[13] The task of evidence identification can serve as a first step for future question answering systems, that can further learn to form abstractive summaries when required based on identifying relevant evidence.

[14] We utilize the HuggingFace implementtation available at https://github.com/huggingface/pytorch-transformers

4.2 Privacy Question Answering

Our goal is to identify evidence within a privacy policy for questions asked by a user. This is framed as an answer sentence selection task, where models identify a set of evidence sentences from all candidate sentences in each policy.

4.2.1 Evaluation Metric

Our evaluation metric for answer-sentence selection is sentence-level F1, implemented similar to (Choi et al., 2018; Rajpurkar et al., 2016). Precision and recall are implemented by measuring the overlap between predicted sentences and sets of gold-reference sentences. We report the average of the maximum F1 from each n−1 subset, in relation to the heldout reference.

4.2.2 Baselines

We describe baselines on this task, including a human performance baseline.

No-Answer Baseline (NA) : Most of the questions we receive are difficult to answer in a legally-sound way on the basis of information present in the privacy policy. We establish a simple baseline to quantify the effect of identifying every question as unanswerable.

Word Count Baseline : To quantify the effect of using simple lexical matching to answer the questions, we retrieve the top candidate policy sentences for each question using a word count baseline (Yang et al., 2015), which counts the number of question words that also appear in a sentence. We include the top 2, 3 and 5 candidates as baselines.

BERT: We implement two BERT-based baselines (Devlin et al., 2019) for evidence identification. First, we train BERT on each query-policy sentence pair as a binary classification task to identify if the sentence is evidence for the question or not (BERT). We also experiment with a two-stage classifier, where we separately train the model on questions only to predict answerability. At inference time, if the answerable classifier predicts the question is answerable, the evidence identification classifier produces a set of candidate sentences (BERT + UNANSWERABLE).

Human Performance: We pick each reference answer provided by an annotator, and compute the F1 with respect to the remaining references, as described in section 4.2.1. Each reference answer is treated as the prediction, and the remaining n-1 answers are treated as the gold reference. The average of the maximum F1 across all reference answers is computed as the human baseline.

5 Results and Discussion

The results of the answerability baselines are presented in Table 5, and on answer sentence selection in Table 6. We observe that BERT exhibits the best performance on a binary answerability identification task. However, most baselines considerably exceed the performance of a majority-class baseline. This suggests considerable information in the question, indicating it's possible answerability within this domain.

Table.6 describes the performance of our baselines on the answer sentence selection task. The No-answer (NA) baseline performs at 28 F1, providing a lower bound on performance at this task. We observe that our best-performing baseline, BERT + UNANSWERABLE achieves an F1 of 39.8. This suggest that BERT is capable of making some progress towards answering questions in this difficult domain, while still leaving considerable headroom for improvement to reach human performance. BERT + UNANSWERABLE performance suggests that incorporating information about answerability can help in this difficult domain. We examine this challenging phenomena of unanswerability further in Section **??**.

5.1 Error Analysis

Disagreements are analyzed based on the OPP-115 categories of each question (Table.7). We compare our best performing BERT variant against the NA model and human performance. We observe significant room for improvement across all categories of questions but especially for first party, third party and data retention categories.

We analyze the performance of our strongest BERT variant, to identify classes of errors and directions for future improvement (Table.8). We observe that a majority of answerability mistakes made by the BERT model are questions which are in fact answerable, but are identified as unanswerable by BERT. We observe that BERT makes 124 such mistakes on the test set. We collect expert judgments on relevance, subjectivity , silence and information about how likely the question is to be answered from the privacy pol-

icy from our experts. We find that most of these mistakes are relevant questions. However many of them were identified as subjective by the annotators, and at least one annotator marked 19 of these questions as having no answer within the privacy policy. However, only 6 of these questions were unexpected or do not usually have an answer in privacy policies. These findings suggest that a more nuanced understanding of answerability might help improve model performance in his challenging domain.

5.2 What makes Questions Unanswerable?

We further ask legal experts to identify potential causes of unanswerability of questions. This analysis has considerable implications. While past work (Rajpurkar et al., 2018) has treated unanswerable questions as homogeneous, a question answering system might wish to have different treatments for different categories of 'unanswerable' questions. The following factors were identified to play a role in unanswerability:

- **Incomprehensibility**: If a question is incomprehensible to the extent that its meaning is not intelligible.
- **Relevance**: Is this question in the scope of what could be answered by reading the privacy policy.
- **Ill-formedness**: Is this question ambiguous or vague. An ambiguous statement will typically contain expressions that can refer to multiple potential explanations, whereas a vague statement carries a concept with an unclear or soft definition.
- **Silence**: Other policies answer this type of question but this one does not.
- **Atypicality**: The question is of a nature such that it is unlikely for any policy policy to have an answer to the question.

Our experts attempt to identify the different 'unanswerable' factors for all 573 such questions in the corpus. 4.18% of the questions were identified as being incomprehensible (for example, 'any difficulties to occupy the privacy assistant'). Amongst the comprehendable questions, 50% were identified as likely to have an answer within the privacy policy, 33.1% were identified as being privacy-related questions but not within the scope of a privacy policy (e.g., *'has Viber had any privacy breaches in the past?'*) and 16.9% of questions were identified as completely out-of-scope (e.g., *"will the app consume much space?'*). In the questions identified as relevant, 32% were ill-formed questions that were phrased by the user in a manner considered vague or ambiguous. Of the questions that were both relevant as well as 'well-formed', 95.7% of the questions were not answered by the policy in question but it was reasonable to expect that a privacy policy would contain an answer. The remaining 4.3% were described as reasonable questions, but of a nature generally not discussed in privacy policies. This suggests that the answerability of questions over privacy policies is a complex issue, and future systems should consider each of these factors when serving user's information seeking intent.

We examine a large-scale dataset of "natural" unanswerable questions (Kwiatkowski et al., 2019) based on real user search engine queries to identify if similar unanswerability factors exist. It is important to note that these questions have previously been filtered, according to a criteria for bad questions defined as "(questions that are) ambiguous, incomprehensible, dependent on clear false presuppositions, opinion-seeking, or not clearly a request for factual information." Annotators made the decision based on the content of the question without viewing the equivalent Wikipedia page. We randomly sample 100 questions from the development set which were identified as unanswerable, and find that 20% of the questions are not questions (e.g., "all I want for christmas is you mariah carey tour"). 12% of questions are unlikely to ever contain an answer on Wikipedia, corresponding closely to our atypicality category. 3% of questions are unlikely to have an answer anywhere (e.g., 'what guides Santa home after he has delivered presents?'). 7% of questions are incomplete or open-ended (e.g., 'the south west wind blows across nigeria between'). 3% of questions have an unresolvable coreference (e.g., 'how do i get to Warsaw Missouri from here'). 4% of questions are vague, and a further 7% have unknown sources of error. 2% still contain false presuppositions (e.g., 'what is the only fruit that does not have seeds?') and the remaining 42% do not have an answer within the document. This reinforces our belief that though they have been understudied in past

work, any question answering system interacting with real users should expect to receive such unanticipated and unanswerable questions.

6 Conclusion

We present PRIVACYQA, the first significant corpus of privacy policy questions and more than 3500 expert annotations of relevant answers. The goal of this work is to promote question-answering research in the specialized privacy domain, where it can have large real-world impact. Strong neural baselines on PRIVACYQA achieve a performance of only 39.8 F1 on this corpus, indicating considerable room for future research. Further, we shed light on several important considerations that affect the *answerability* of questions. We hope this contribution leads to multidisciplinary efforts to precisely understand user intent and reconcile it with information in policy documents, from both the privacy and NLP communities.

Acknowledgements

This research was supported in part by grants from the National Science Foundation Secure and Trustworthy Computing program (CNS-1330596, CNS-1330214, CNS-15-13957, CNS-1801316, CNS-1914486, CNS-1914444) and a DARPA Brandeis grant on Personalized Privacy Assistants (FA8750-15-2-0277). The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation. 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, DARPA, or the US Government. The authors would like to extend their gratitude to Elias Wright, Gian Mascioli, Kiara Pillay, Harrison Kay, Eliel Talo, Alexander Fagella and N. Cameron Russell for providing their valuable expertise and insight to this effort. The authors are also grateful to Eduard Hovy, Lorrie Cranor, Florian Schaub, Joel Reidenberg, Aditya Potukuchi and Igor Shalyminov for helpful discussions related to this work, and to the three anonymous reviewers of this draft for their constructive feedback. Finally, the authors would like to thank all crowdworkers who consented to participate in this study.

References

Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The Guardian*, 17.

Fred H Cate. 2010. The limits of notice and choice. *IEEE Security & Privacy*, 8(2).

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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

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

SS Deeb. 2005. The molecular basis of variation in human color vision. *Clinical genetics*, 67(5):369–377.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking svm and deep convolutional neural network. *arXiv preprint arXiv:1703.05320*.

Federal Trade Commission. 1998. Privacy online: A report to congress. *Washington, DC, June*, pages 10–11.

Minwei Feng, Bing Xiang, Michael R Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Anette Frank, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg, and Ulrich Schäfer. 2007. Question answering from structured knowledge sources. *Journal of Applied Logic*, 5(1):20–48.

Matthias Gallé, Athena Christofi, and Hady Elsahar. 2019. The case for a gdpr-specific annotated dataset of privacy policies. *AAAI Symposium on Privacy-Enhancing AI and HLT Technologies*.

Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 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)*, pages 321–340.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. *arXiv preprint arXiv:1802.02561*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Priyank Jain, Manasi Gyanchandani, and Nilay Khare. 2016. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3(1):25.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. 2009. A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 4. ACM.

Mi-Young Kim, Ying Xu, and Randy Goebel. 2015. Applying a convolutional neural network to legal question answering. In *JSAI International Symposium on Artificial Intelligence*, pages 282–294. Springer.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Pedro Giovanni Leon, Ashwini Rao, Florian Schaub, Abigail Marsh, Lorrie Faith Cranor, and Norman Sadeh. 2015. Privacy and behavioral advertising: Towards meeting users' preferences. In *Symposium on usable privacy and security (SOUPS)*.

Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2016. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *2016 AAAI Fall Symposium Series*.

Yi-Hung Liu, Yen-Liang Chen, and Wu-Liang Ho. 2015. Predicting associated statutes for legal problems. *Information Processing & Management*, 51(1):194–211.

Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *ISJLP*, 4:543.

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

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

Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. 2009. Nlp for shallow question answering of legal documents using graphs. *Computational Linguistics and Intelligent Text Processing*, pages 498–508.

Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2774–2779, Copenhagen, Denmark. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

OCDE OECD. 2004. The oecd principles of corporate governance. *Contaduría y Administración*, (216).

Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2017. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web*, (Preprint):1–19.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas. Association for Computational Linguistics.

P Howard Patrick. 1980. Privacy restrictions on transnational data flows: A comparison of the council of europe draft convention and oecd guidelines. *Jurimetrics J.*, 21:405.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Paulo Quaresma and Irene Pimenta Rodrigues. 2005. A question answer system for legal information retrieval. In *JURIX*, pages 91–100.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ*, 30:39.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17.

Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Which apps have privacy policies? In *Annual Privacy Forum*, pages 3–23. Springer.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 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)*, pages 464–473, Berlin, Germany. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

FTC US Federal Trade Commission et al. 2012. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. *FTC Report*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.

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

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, 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, pages 1330–1340.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018.