The Creation and Analysis of a Website Privacy Policy Corpus

Shomir Wilson[1], Florian Schaub[1], Aswarth Abhilash Dara[1], Frederick Liu[1], Sushain Cherivirala[1], Pedro Giovanni Leon[2], Mads Schaarup Andersen[1], Sebastian Zimmeck[3], Kanthashree Mysore Sathyendra[1], N. Cameron Russell[4], Thomas B. Norton[4], Eduard Hovy[1], Joel Reidenberg[4] and Norman Sadeh[1]

[1]Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, USA
[2]Stanford University, Center for Internet and Society, Stanford, CA, USA
[3]Columbia University, Department of Computer Science, New York, NY, USA
[4]Fordham University, Law School, New York, NY, USA
shomir@cs.cmu.edu, sadeh@cs.cmu.edu

Abstract

Website privacy policies are often ignored by Internet users, because these documents tend to be long and difficult to understand. However, the significance of privacy policies greatly exceeds the attention paid to them: these documents are binding legal agreements between website operators and their users, and their opaqueness is a challenge not only to Internet users but also to policy regulators. One proposed alternative to the status quo is to automate or semi-automate the extraction of salient details from privacy policy text, using a combination of crowdsourcing, natural language processing, and machine learning. However, there has been a relative dearth of datasets appropriate for identifying data practices in privacy policies. To remedy this problem, we introduce a corpus of 115 privacy policies (267K words) with manual annotations for 23K fine-grained data practices. We describe the process of using skilled annotators and a purpose-built annotation tool to produce the data. We provide findings based on a census of the annotations and show results toward automating the annotation procedure. Finally, we describe challenges and opportunities for the research community to use this corpus to advance research in both privacy and language technologies.

1 Introduction

Privacy policies written in natural language are a nearly pervasive feature of websites and mobile applications. The "notice and choice" legal regimes of many countries require that website operators post a notice of how they gather and process users' information. In theory, users then choose whether to accept those practices or to abstain from using the website or service. In practice, however, the average Internet user struggles to understand the contents of privacy policies (McDonald and Cranor, 2008) and generally does not read them (Federal Trade Commission, 2012; President's Concil of Advisors on Science and Technology, 2014). This disconnect between Internet users and the data practices that affect them has led to the assessment that the notice and choice model is ineffective in the status quo (Reidenberg et al., 2015b; Cate, 2010).

Thus, an opening exists for language technologies to help "bridge the gap" between privacy policies in their current form and representations that serve the needs of Internet users. Such a bridge would also serve unmet needs of policy regulators, who do not have the means to assess privacy policies in large numbers. Legal text is a familiar domain for natural language processing, and the legal community has demonstrated some reciprocal interest (Mahler, 2015). However, the scale of the problem and its significance—i.e., to virtually any Internet user, as well as to website operators and policy regulators—distinguishes it and provides immense motivation (Sadeh et al., 2013).

To this end, we introduce a corpus of 115 website privacy policies annotated with detailed information about the data practices that they describe.[1] This information consists of 23K data practices, 128K practice attributes, and 103K annotated text spans, all produced by skilled anno-

[1]The dataset is available for download at www.usableprivacy.org/data.

tators. To the best of our knowledge, this is the first large-scale effort to annotate privacy policies at such a fine level of detail. It exceeds prior efforts to annotate sentence-level fragments of policy text (Breaux and Schaub, 2014), answer simple overarching questions about privacy policy contents (Wilson et al., 2016; Zimmeck and Bellovin, 2014), or analyze the readability of privacy policies (Massey et al., 2013). We further present analysis that demonstrates the richness of the corpus and the feasibility of partly automating the annotation of privacy policies.

The remainder of this paper is structured as follows. We discuss related work and contextualize the corpus we have created in Section 2. In Section 3 we describe the creation of the corpus, including the collection of a diverse set of policies and the creation of a privacy policy annotation tool. Section 4 presents analysis that illustrates the diversity and complexity of the corpus, and Section 5 shows results on the prediction of policy structure. Finally, in Section 6 we describe some promising avenues for future work.

2 Related Work

Prior attempts on analyzing privacy policies focused largely on manually assessing their usability (Jensen and Potts, 2004) or compliance with self-regulatory requirements (Hoke et al., 2015). Breaux et al. proposed a description logic to analyze and reason about data sharing properties in privacy policies (2013), but rely on a small set of manually annotated privacy policies to instantiate their language. Automated assessments have largely focused on readability scores (Massey et al., 2013; Meiselwitz, 2013; Ermakova et al., 2015). Cranor et al. leveraged the standardized format of privacy notices in the U.S. financial industry to automatically analyze privacy polices of financial institutions (2013). However, in spite of notable efforts such as P3P (Wenning et al., 2006), the majority of privacy policies are unstructured and do not follow standardized formats.

Costante et al. (2012) proposed a supervised learning approach to determine which data practice categories are covered in a privacy policy. Rule-based extraction techniques have been proposed to extract some of a website's data collection practices from its privacy policy (Costante et al., 2013) or to answer certain binary questions about a privacy policy (Zimmeck and Bellovin, 2014). Other approaches leverage topic modeling (Chundi and Subramaniam, 2014; Stamey and Rossi, 2009) or sequence alignment techniques (Liu et al., 2014; Ramanath et al., 2014) to analyze privacy policies or identify similar policy sections and paragraphs. However, the complexity and vagueness of privacy policies makes it difficult to automatically extract complex data practices from privacy policies without substantial gold standard data.

Crowdsourcing has been proposed as a potential approach to obtain annotations for privacy policies (Sadeh et al., 2013; Breaux and Schaub, 2014; Wilson et al., 2016). However, crowdworkers are not trained in understanding and interpreting legal documents, which may result in interpretation discrepancies compared to experts (Reidenberg et al., 2015a). Our policy annotation tool shares some common features with GATE (Bontcheva et al., 2013), although the interface for our tool is simpler to fit the specific requirements of the task.

Few prior efforts, aside from those we cite above, have applied natural language processing to privacy policies or other legal documents purported for the general public to regularly read. More generally, legal text has a history of attention from natural language processing (Bach et al., 2013; Galgani et al., 2012; Francesconi et al., 2010) and from artificial intelligence (Sartor and Rotolo, 2013; Bench-Capon et al., 2012). Classifying legal text into categories has received some interest (Šavelka and Ashley, 2015; Mickevicius et al., 2015), as well as making the contents of legal texts more accessible (Boella et al., 2015; Curtotti and McCreath, 2013).

Compared to prior efforts, our data set is notable for its combination of size, input from experts (for the label scheme) and skilled annotators (for the annotation procedure), and fine-grained detail.

3 Corpus Creation and Structure

In this section we describe our procedure for selecting a diverse set of privacy policies, our annotation scheme, how we obtained annotations, and the structure of the corpus.

3.1 Privacy Policy Selection

Privacy policies vary in length, complexity, legal sophistication, and coverage of services. For instance, privacy policies of large companies may cover multiple services, websites, apps, and even physical stores; such policies are often crafted by legal teams and frequently updated. Privacy poli-

cies of smaller or less popular companies may have narrower focus or vary in employed language, and they may be updated less frequently.

To reflect this diversity, we used a two-step process for policy selection: (1) relevance-based website pre-selection and (2) sector-based sub-sampling. First, we monitored Google Trends (Google, 2015) for one month (May 2015) to collect the top five search queries for each trend. Then, for each query we retrieved the first five websites listed on each of the first 10 pages of results. This process produced a diverse sample of 1,799 unique websites.

Second, we sub-sampled from this website dataset according to DMOZ.org's top-level website sectors.[2] More specifically, we organized the dataset into 15 sectors (e.g., Arts, Shopping, Business, News). We excluded the "World" sector and limited the "Regional" sector to the "U.S." subsector in order to ensure that all privacy policies in our corpus are subject to the same legal and regulatory requirements. We ranked the websites in each sector according to their frequency in the retrieved search results. Then we selected eight websites from each sector by randomly chosing two websites from each rank quartile.

For each selected website, we manually verified that it had an English-language privacy policy and that it pertained to a US company (based on contact information and WHOIS entry) before downloading its privacy policy. Excluded websites were replaced with random re-draws from the same sector rank quartile. Some privacy policies covered more than one of the selected websites (e.g., the Disney privacy policy covered disney.go.com and espn.go.com), resulting in a final dataset of 115 privacy policies across 15 sectors.

3.2 Annotation Scheme and Process

We developed a policy annotation scheme to capture the data practices specified by privacy policies. To ensure the scheme reflected actual policy contents, development occurred as an iterative refinement process, in which a small group of domain experts (privacy experts, public policy experts, and legal scholars) identified different data practice categories and their descriptive attributes from multiple privacy policies. The annotation scheme was then applied to additional policies and refined over multiple iterations during discussions among the experts.

The final annotation scheme consists of ten data practice categories:

1. *First Party Collection/Use*: how and why a service provider collects user information.
2. *Third Party Sharing/Collection*: how user information may be shared with or collected by third parties.
3. *User Choice/Control*: choices and control options available to users.
4. *User Access, Edit, & Deletion*: if and how users may access, edit, or delete their information.
5. *Data Retention*: how long user information is stored.
6. *Data Security*: how user information is protected.
7. *Policy Change*: if and how users will be informed about changes to the privacy policy.
8. *Do Not Track*: if and how Do Not Track signals[3] for online tracking and advertising are honored.
9. *International & Specific Audiences*: practices that pertain only to a specific group of users (e.g., children, Europeans, or California residents).
10. *Other*: additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.

An individual *data practice* belongs to one of the ten categories above, and it is articulated by a category-specific set of *attributes*. For example, a User Choice/Control data practice is associated with four mandatory attributes (*Choice Type*, *Choice Scope*, *Personal Information Type*, *Purpose*) and one optional attribute (*User Type*). The annotation scheme defines a set of potential values for each attribute. To ground the data practice in the policy text, each attribute also may be associated with a *text span* in the privacy policy.

The set of mandatory and optional attributes reflects the potential level of specificity with which a data practice of a given category may be described. Optional attributes are less common, while mandatory attributes are necessary to represent a data practice. However, privacy policies are often vague or ambiguous on many of these attributes. Therefore, a valid value for each attribute is *Unspecified*, allowing annotators to express an absence of information.

[2] The DMOZ.org website sectors are notable for their use by Alexa.com.

[3] www.w3.org/2011/tracking-protection

Documents	115
Words	266,713
Annotated Data Practices	23,194
Annotated Attributes	128,347
Annotated Text Spans	102,576
Annotators Per Document	3
Annotators Total	10

Table 1: Totalized statistics on the corpus.

We developed a web-based annotation tool, shown in Figure 1, for skilled annotators to apply our annotation scheme to the selected privacy policies.[4] In preparation, privacy policies were divided into paragraph-length *segments* for annotators to read in the tool, one at a time in sequence. For each segment, an annotator may label zero or more data practices from each category. To create a data practice, an annotator first selects a practice category and then specifies values and text spans for each of its attributes. Annotators can see a list of data practices they have created for a segment and selectively duplicate and edit them to annotate practices that differ only slightly, though we omit these features from the figure for brevity.

4 Composition of the OPP-115 Corpus

The annotation process produced a nuanced and diverse dataset, which we describe in detail below. We name the dataset the *OPP-115 Corpus* (Online Privacy Policies, set of 115) for convenience.

4.1 Policy Contents

Table 1 shows some descriptive statistics for the corpus as a whole. Each privacy policy was read by three skilled annotators, who worked independently, and a total of ten annotators participated in the process. All annotators were law students and were compensated for their work at rates appropriate for student employees at their respective universities. They required a mean of 72 minutes per policy, though this number is slightly inflated by outliers when they stepped away from in-progress sessions for extended periods of time. The annotators produced a total of 23K data practices, although this number contains some redundancies between annotators' efforts.[5] In aggregate, these data practices are associated with 128K values for attributes and 103K selected spans of policy text. Note that the annotation tool required the selection of a text span for mandatory attributes, but did not require a text-based justification for optional attributes or attributes marked as "Unspecified".

The corpus allows us to investigate the composition of typical privacy policies in terms of data practices. Privacy policies are known for their length and complexity, but those notions do not necessarily entail a density of pertinent information. Table 2 shows the pre-consolidation quantities of practices that we collected in each of the ten annotation categories, along with the mean and median counts of practices per privacy policy. Intuitively, First Party Collection/Use and Third Party Sharing/Collection dominated the rankings by frequency: the collection, usage, and sharing of user data are the primary concerns that compel the production of privacy policies. Data practices in the Other category, while frequent, were mostly statements that were ostensibly not about user data; 57% were introductory, contact, or generic information. Means were above medians for all categories, reflecting rightward skews for all the distributions.

Table 2 also contains statistics on segment-level category coverage and annotator agreement. Here, *coverage* is meant in an *ipso facto* sense: a practice category covers a policy segment if two of three annotators each identified at least one practice from that category in the segment text. Differences in the category rankings by frequency and by coverage reveal that practices in some categories are less tightly clustered than others. In particular, Data Retention is the second rarest practice category but ranks fourth by segment coverage. Since Kappa is applied here to an artificial task (annotators were not asked to label entire segments) the common conventions for its interpretation (Carletta, 1996; Viera and Garrett, 2005) are not directly applicable. However, Do Not Track and International and Specific Audiences remain standout categories with the greatest segment-level agreement. We hypothesize that these two categories have the most easily recognizable cues for annotation. Do Not Track practices, for example, are associated with the eponymous phrase.

Finally, the pre-consolidation mean and median

[4] Our experts for the annotation scheme development and our skilled annotators were mutually exclusive groups.

[5] We describe a method to *consolidate* annotations (i.e., to eliminate redundancies between annotators' data) in Section 4.2. Here, we analyze policy contents pre-consolidation to avoid propagating the effects of nontrivial assumptions necessary for consolidation.

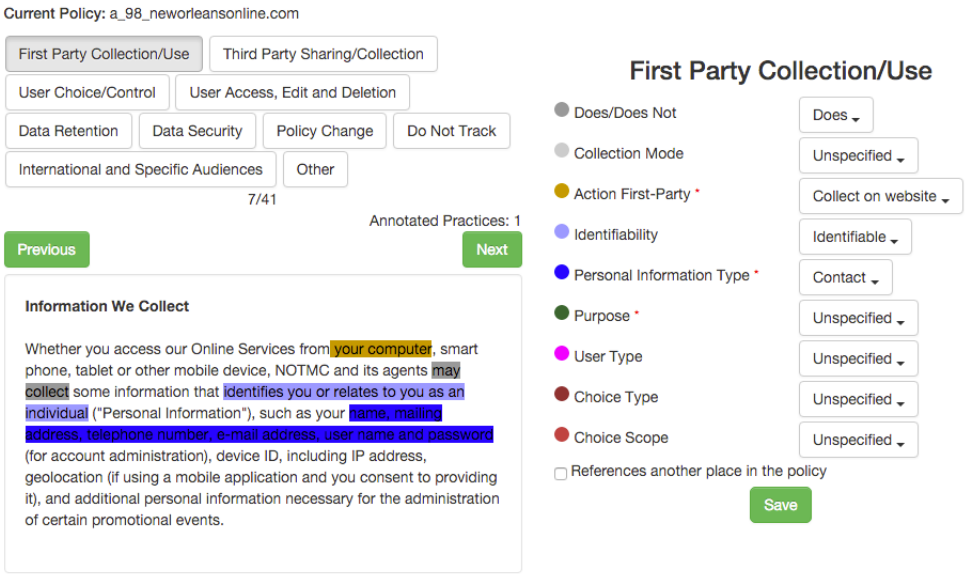

Figure 1: Web-based tool for our skilled annotators of privacy policies.

Category	Freq.	Mean	Median	Coverage	Fleiss' Kappa
First Party Collection/Use	8,956	78	74	.27	.76
Third Party Sharing/Collection	5,230	45	39	.21	.76
Other	3,551	31	25	.24	.49
User Choice/Control	1,791	16	13	.08	.61
Data Security	1,009	9	7	.05	.67
International and Specific Audiences	941	8	6	.07	.87
User Access, Edit and Deletion	747	6	5	.03	.74
Policy Change	550	5	4	.03	.73
Data Retention	370	3	2	.20	.55
Do Not Track	90	1	0	.01	.91

Table 2: By-category descriptive statistics for the data practices in the corpus. These statistics are calculated prior to consolidating multiple annotators' work. Means and medians are calculated across the population of policies in the corpus. Coverage and Kappa are calculated in terms of by-segment contents.

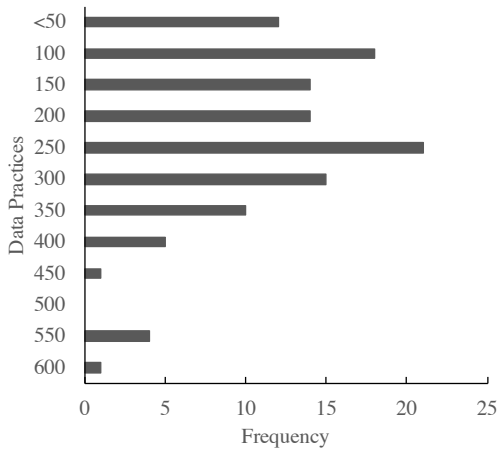

Figure 2: Distribution of data practices per policy.

quantities of data practices per policy were 202 and 200, respectively. These do not correspond to columnar totals that may be calculated from Table 2 because the categories were not equally distributed among the privacy policies. Figure 2 shows the distribution of quantities of data practices per policy. The distribution exhibits a skew toward larger numbers of data practices per policy. Importantly, differences in the number of data practices should not be interpreted as varying levels of data protection or privacy. A privacy policy that contains many data practices may exhibit substantial redundancy among them, and a privacy policy with relatively few data practices could merely be concise. In either case, the data practices may be responsive to users concerns or at odds with them.

4.2 Consolidating Annotators' Work

In this section we discuss the problem of *consolidation*, or merging data practices from multiple annotators if those practices refer to the same underlying practice expressed by the text. The ambiguity and vagueness of privacy policies (Reidenberg et al., 2016) and the sophistication of the annotation scheme are natural limitations on annotator agreement. With that in mind, we present a consolidation procedure to collapse redundant annotations with the proviso that practices labeled by only one or two skilled annotators also have substantial value and merit retention.

First, we institute some basic requirements about locality and topicality to determine which data practices are eligible for consolidation. Given a segment, if annotators $A_1, ..., A_n$ (for $n = 2$ or $n = 3$ in our dataset) respectively produce sets of data practices $P_1, ..., P_n$, then a selection of data practices $p_1 \in P_1, ..., p_n \in P_n$ is eligible to be consolidated into a single data practice only if all of them belong to the same category. Additionally, three implicit assumptions in this requirement are that (1) at least two annotators contribute practices to a consolidation set, (2) all the practices are located in the same policy segment, and (3) each practice must belong to a unique annotator.

For each segment we create an exhaustive list of eligible combinations of data practices to consolidate, score and rank each combination using a method detailed below, prune the list with a score threshold, and finally perform consolidations in order of ranking until no further consolidations are possible. Consolidation sets containing three annotators' practices are considered prior to sets containing practices from only two annotators. The data practices in a chosen consolidation set are removed and replaced by a single "master" data practice. To do this, it is necessary to merge sets of values and sets of text spans respectively associated with each attribute. Sets of values are merged using a majority vote if possible and set to Unspecified if otherwise; the latter case occurs in approximately a third of all mergers. Sets of text spans are merged with a strong bias toward recall, by creating a new text span that begins and ends with respectively the first and last indexes in the set.

Our scoring method is based on the summative overlap between the sets of text spans associated with attributes of data practices, with normalization to account for longer spans. Since the text spans connect data practices to the policy text, we use their overlaps as evidence that two annotators' data practices refer to the same underlying practice in the text. Thus, a score for two data practices that are associated with roughly the same policy text is relatively high, and a score for two data practices that are associated with different text is low.

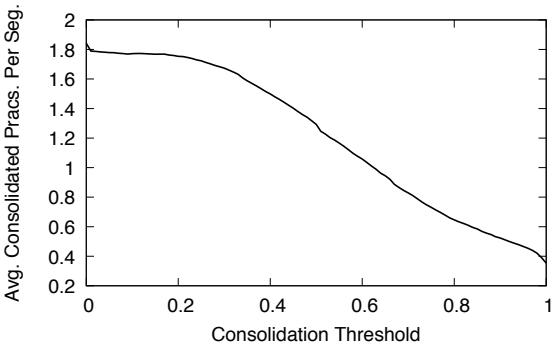

Figure 3: Consolidation threshold value versus the average number of data practices per segment produced by consolidation.

Figure 3 shows the effect of the consolidation threshold on the average number of practices produced by consolidation per policy segment (i.e., excluding those original data practices that were retained because they were not subject to consolidation). Past a threshold value of approximately 0.2, the number of practices steadily decreases. Notably the average number of practices produced by consolidation is substantially less than the average practices per annotator per segment (2.04) at any point on the curve, indicating a relative lack of agreement between annotators in terms of text span selections. As part of the corpus, we release consolidated datasets at threshold values of 0.5, 0.75, and 1.

4.3 Data Exploration Website

The data practice annotations are difficult for human readers to interpret without visual connections to the policy text. To help researchers, policy regulators, and the general public understand the structure and utility of the data set, we created a data exploration website[6] that visually integrates the data practice annotations with the texts of privacy policies. Site interactivity allows users to search for websites in the dataset or browse by DMOZ sectors.

The website also allows users to compare privacy policies by categorical structure, data prac-

[6] explore.usableprivacy.org

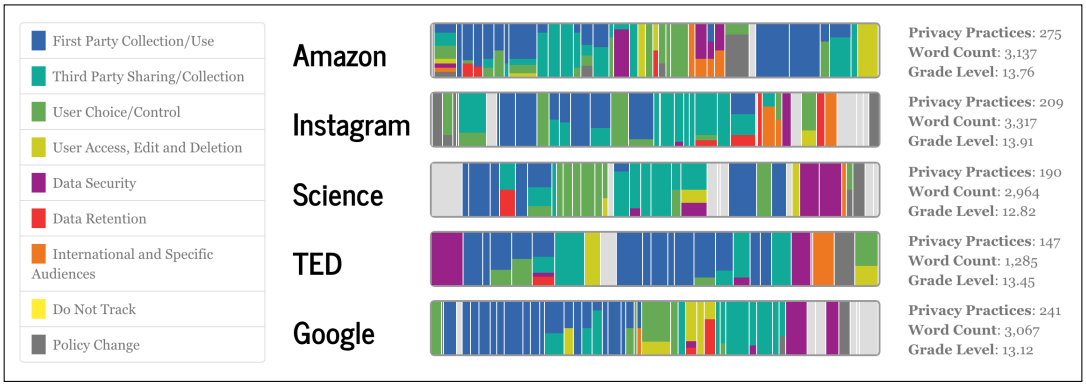

Figure 4: Comparing five policies on the data exploration website.

tice quantities, and reading level.[7] Figure 4 shows a sample comparison between five websites. Each policy's segments are depicted in order from left to right. Segments are colored according to the practice categories that annotators labeled within them. Qualitative patterns are discernible; for example, several of these policies have large blocks of First Party Collection/Use toward the beginning or large blocks of Third Party Sharing/Collection further inward. We discuss exploiting such recurring structures in the next section.

5 Prediction of Policy Structure

The current human annotation procedure is impractical for covering the entire Internet or accounting for changes in privacy policies. This raises the question of whether the process can be partly automated. In this section we describe our experiments to automatically assign category labels to policy segments, which would enable the simplification of the annotation task.

5.1 Experiments

Our dataset consisted of 3,792 segments from 115 privacy policies. We represented the text of each segment as a dense vector using Paragraph2Vec (Le and Mikolov, 2014) and the GENSIM toolkit (Řehůřek and Sojka, 2010). This approach exploited semantic similarities between words in the vocabulary of privacy policies, acknowledging that the vocabulary in this domain is specialized but not completely standardized. We assigned each policy segment a binary vector of category-specific labels, with each element in the vector corresponding to the presence or absence of a data practice category in the segment. We considered a vector with twelve elements, with nine of them coming from existing practice categories (all except Other). The remaining three came from elevating three attributes of Other to category status: *Introductory/Generic*, *Practice Not Covered*, and *Privacy Contact Information*. We created gold standard data for this problem using a simplified consolidation approach: if two or more annotators agreed that a category is present in a segment, then we labeled that segment with the category.

To predict the category labels of privacy policy segments, we tried three approaches. Two were logistic regression and SVM models, for which we treated this as a multi-class classification problem. Since 2^{12} unique category vectors exist, we trimmed the label space to only those that occur in the training set. The third was a sequence labeling approach inspired by prior work to apply hidden Markov models (HMMs) to privacy policy text (Ramanath et al., 2014). Our work differs from this prior work by using labels from an annotation scheme constructed by privacy experts rather than topics developed from an unsupervised method. Additionally, in our formulation, each hidden state corresponds to one of the unique binary vectors that represent classes of category combinations in the training data. The HMM's transition probabilities capture the tendency of privacy policy authors to organize topics (i.e., practice categories in our annotation scheme) in similar sequences. Since each segment is represented by a unique real-valued vector from Paragraph2Vec, it was not possible to directly obtain an emission probability distribution from the training data. Therefore, we ran the K-Means++ algorithm using the scikit-learn toolkit (Pedregosa et al., 2011) on the segment vector representations and assigned each segment to a cluster. The emission probability distribution then captured the tendencies of a given class and generated the segment

[7] `explore.usableprivacy.org/compare`

that is represented as a cluster. These two distributions are estimated empirically from the training data, and we used Viterbi decoding to obtain the best labeling sequence during the prediction.

5.2 Results

We split the set of 115 policies into subsets of 75 for training and 40 for testing. The number of clusters in the HMM approach[8] is set to 100 and the results are shown in Table 3 as means across 10 runs. The standard deviations for these performance figures are generally between 0.01 and 0.05; the one exception is Do Not Track (the least frequent category) with a standard deviation of 0.2. As the table shows, although the HMM does not reach the same performance as SVM, it performs similarly to logistic regression and meets or exceeds its F1-score for five categories.

We interpret the strength of the SVM as indicating the strong potential to partly automate the policy labeling procedure, especially for two categories: First Party Collection/Use (a standout performance and the category for which the most labeled data exists) and Do Not Track (a perfect performance, likely due to the limited vocabulary used to describe practices in this category). Additionally, while the HMM did not perform as well overall, we note that its micro-average F1 was a slight improvement over logistic regression. With relatively little data to train this HMM, we expect that the accumulation of more labeled instances can improve its performance substantially.

6 Future Directions

The OPP-115 Corpus enables research in several directions of interest to natural language processing and usable privacy. We sketch some opportunites for future work below.

A central challenge for this research is the scalability of policy annotation. Although it was necessary to annotate the first 115 policies manually, to ensure the annotations were responsive to the annotation scheme, a less labor-intensive approach will be required for large-scale Web coverage. The OPP-115 Corpus is a valuable dataset for this move toward automated methods. Additionally, a strong potential exists for a combination of automated annotation of coarse information and human annotation of finer details. For example, automated category labeling of policy segments is feasible, as demonstrated in Section 5. Asking a human to label practices in a single category would be a reduction in effort, especially if they are shown text that is relevant to the category. Crowdsourcing also becomes a possibility when the complexity of the task is reduced.

An ambitious goal will be to eliminate human annotators altogether. Our preliminary analysis has shown that the policy vocabularies associated with certain annotations are very distinctive (e.g., the Do Not Track category or financial information as a data type, for example), lending themselves to automatic identification. By producing confidence ratings alongside data practice predictions, an automated system could mitigate its shortcomings.

Separately, data practices must be presented to Internet users in a way that is responsive to their concerns. Text summarization is a possibility, using the annotations as a guide for important details to retain. Internet users have already demonstrated limited patience with text-based privacy policies, which adds a nuance to this challenge and suggests the need for a combination of text and pictorial representations (or chiefly pictorial representations) to communicate data practices (Schaub et al., 2015).

Additional questions of interest include:

- *How can the data practice annotations for a policy be combined into a cohesive interpretation?* The relationships between data practices are not straightforward. Vagueness, contradictions, and unclear scope are all problems for constructing a knowledge base of them.
- *How can the balance between human and automated methods for annotation be optimized?* The model for the ideal combination is subject to parameters such as the availability of resources and the necessary level of confidence for annotations.
- *How can sectoral norms and outliers be identified automatically?* A bank website that collects users' health information, for example, deserves scrutiny. It seems appropriate to address this question with clustering techniques, using features from the data practices and from the policy text.

[8] We tuned the parameters of the HMM approach and SVM after performing a five-fold cross validation on the training data.

	LR			SVM			HMM		
Category	P	R	F	P	R	F	P	R	F
First Party Collection/Use	0.73	0.67	0.70	0.76	0.73	0.75	0.69	0.76	0.72
Third Party Sharing/Collection	0.64	0.63	0.63	0.67	0.73	0.70	0.63	0.61	0.62
User Choice/Control	0.45	0.62	0.52	0.65	0.58	0.61	0.47	0.33	0.39
Introductory/Generic*	0.51	0.50	0.50	0.58	0.49	0.53	0.54	0.49	0.51
Data Security	0.48	0.75	0.59	0.66	0.67	0.67	0.67	0.53	0.59
Internat'l and Specific Audiences	0.49	0.69	0.57	0.70	0.70	0.70	0.67	0.66	0.66
Privacy Contact Information*	0.34	0.72	0.46	0.60	0.68	0.64	0.48	0.59	0.53
User Access, Edit, and Deletion	0.47	0.71	0.57	0.67	0.56	0.61	0.48	0.42	0.45
Practice Not Covered*	0.20	0.47	0.28	0.19	0.26	0.22	0.15	0.12	0.13
Policy Change	0.59	0.83	0.69	0.66	0.88	0.75	0.52	0.68	0.59
Data Retention	0.10	0.35	0.16	0.12	0.12	0.12	0.08	0.12	0.09
Do Not Track	0.45	1.0	0.62	1.0	1.0	1.0	0.45	0.40	0.41
Micro-Average	0.53	0.65	0.58	0.66	0.66	0.66	0.60	0.59	0.60

Table 3: Precision/Recall/F1 for the three models. The three starred categories resulted from the decomposition of the original Other category, which is excluded here. Categories are ordered in this table in descending order by frequency in the dataset.

7 Conclusion

We have described the motivation, creation, and analysis of a unique corpus of 115 privacy policies and 23K fine-grained data practice annotations, and we have demonstrated the feasibility of partly automating the annotation process. The annotations reveal the structure and complexity of these documents, which Internet users are expected to understand and accept. This corpus should serve as a resource for language technologies research to help Internet users understand the privacy practices of businesses and other entities that they interact with online.

Acknowledgements

This work is funded by the National Science Foundation under grants CNS-1330596 and CNS-1330214. The authors would like to acknowledge the law students at Fordham University and the University of Pittsburgh who worked as annotators to make this corpus possible. The authors also wish to acknowledge all members of the Usable Privacy Policy Project (`www.usableprivacy.org`) for their contributions.

References

Ngo Xuan Bach, Nguyen Le Minh, Tran Thi Oanh, and Akira Shimazu. 2013. A two-phase framework for learning logical structures of paragraphs in legal articles. *ACM Transactions on Asian Language Information Processing (TALIP)*, 12(1):3:1–3:32, March.

Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G Conrad, Enrico Francesconi, et al. 2012. A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law. *Artificial Intelligence and Law*, 20(3):215–319.

Guido Boella, Luigi Di Caro, Michele Graziadei, Loredana Cupi, Carlo Emilio Salaroglio, Llio Humphreys, Hristo Konstantinov, Kornel Marko, Livio Robaldo, Claudio Ruffini, et al. 2015. Linking legal open data: Breaking the accessibility and language barrier in European legislation and case law. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 171–175. ACM.

Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. GATE teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation*, 47(4):1007–1029.

Travis D. Breaux and Florian Schaub. 2014. Scaling requirements extraction to the crowd. In *RE'14: Proceedings of the 22nd IEEE International Requirements Engineering Conference*, RE' 14, pages 163–172, Washington, DC, USA, August. IEEE Society Press.

Travis D. Breaux, Hanan Hibshi, and Ashwini Rao. 2013. Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. *Requirements Engineering*, 19(3):281–307, September.

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

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

Parvathi Chundi and Pranav M. Subramaniam. 2014. An approach to analyze web privacy policy documents. In *KDD Workshop on Data Mining for Social Good*.

Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, WPES '12, pages 91–96. ACM.

Elisa Costante, Jerry den Hartog, and Milan Petković. 2013. What websites know about you: Privacy policy analysis using information extraction. In Roberto Di Pietro, Javier Herranz, Ernesto Damiani, and Radu State, editors, *Data Privacy Management and Autonomous Spontaneous Security*, volume 7731 of *Lecture Notes in Computer Science*, pages 146–159. Springer.

Lorrie Faith Cranor, Kelly Idouchi, Pedro Giovanni Leon, Manya Sleeper, and Blase Ur. 2013. Are they actually any different? Comparing thousands of financial institutions' privacy practices. In *Workshop on the Economics of Information Security*, WEIS '13, June.

Michael Curtotti and Eric McCreath. 2013. A right to access implies a right to know: An open online platform for research on the readability of law. *J. Open Access L.*, 1:1–56.

Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. 2015. Readability of privacy policies of healthcare websites. In *12. Internationale Tagung Wirtschaftsinformatik (Wirtschaftsinformatik 2015)*.

Federal Trade Commission. 2012. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. `http://www.ftc.gov/reports`. Last accessed: June 22, 2016.

Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. 2010. *Semantic processing of legal texts: Where the language of law meets the law of language*, volume 6036 of *Lecture Notes in Artificial Intelligence*. Springer.

Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Combining different summarization techniques for legal text. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, pages 115–123. ACL.

Google. 2015. Google Trends. `https://www.google.com/trends/`. Last accessed: June 22, 2016.

Candice Hoke, Lorrie Cranor, Pedro Leon, and Alyssa Au. 2015. Are they worth reading? An in-depth analysis of online tracker's privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 11(2):325–404, April.

Carlos Jensen and Colin Potts. 2004. Privacy policies as decision-making tools: An evaluation of online privacy notices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 471–478. ACM.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A. Smith. 2014. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, COLING '14, pages 884–894. ANLP, August.

Lars Mahler. 2015. What is NLP and why should lawyers care? `http://www.lawpracticetoday.org/article/nlp-lawyers/`, February. Last accessed: June 22, 2016.

Aaron K. Massey, Jacob Eisenstein, Annie I. Antón, and Peter P. Swire. 2013. Automated text mining for requirements analysis of policy documents. In *21st IEEE International Requirements Engineering Conference*, RE '13, pages 4–13. IEEE.

Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *I/S: J Law & Policy Info. Soc.*, 4(3):540–561.

Gabriele Meiselwitz. 2013. Readability assessment of policies and procedures of social networking sites. In A. Ant Ozok and Panayiotis Zaphiris, editors, *5th International conference, OCSC 2013, Held as Part of HCI International 2013*, Lecture Notes in Computer Science, pages 67–75. Springer, January.

Vytautas Mickevicius, Tomas Krilavicius, and Vaidas Morkevicius. 2015. Classification of short legal Lithuanian texts. In *The 5th Workshop on Balto-Slavic Natural Language Processing*, BSNLP '15, pages 106–111. ACM.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

President's Concil of Advisors on Science and Technology. 2014. Big data and privacy: a technological perspective. Report to the President, Executive Office of the President, May.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A. Smith. 2014. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, ACL '14, pages 605–610. ACL, June.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA, May.

Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T. Graves, Fei Liu, Aleecia McDonald, Thomas B. Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. 2015a. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30(1):39–88.

Joel R. Reidenberg, N. Cameron Russell, Alexander J. Callen, Sophia Qasir, and Thomas B. Norton. 2015b. Privacy harms and the effectiveness of the notice and choice framework. *I/S: J Law & Policy Info. Soc.*, 11(2).

Joel R. Reidenberg, Jaspreet Bhatia, Travis Breaux, and Thomas B. Norton. 2016. Automated comparisons of ambiguity in privacy policies and the impact of regulation. *Social Science Research Network Working Paper Series*, January.

Norman Sadeh, Alessandro Acquisti, Travis D. Breaux, Lorrie Faith Cranor, Aleecia M. Mcdonald, Joel R. Reidenberg, Noah A. Smith, Fei Liu, N. Cameron Russell, Florian Schaub, and Shomir Wilson. 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.

Giovanni Sartor and Antonino Rotolo. 2013. AI and law. In *Agreement Technologies*, pages 199–207. Springer.

Jaromír Šavelka and Kevin D Ashley. 2015. Transfer of predictive models for classification of statutory texts in multi-jurisdictional settings. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 216–220. ACM.

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 '15, pages 1–17. USENIX Association, July.

John W. Stamey and Ryan A. Rossi. 2009. Automatically identifying relations in privacy policies. In *27th ACM International Conference on Design of Communication*, SIGDOC '09, pages 233–238. ACM.

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37(5):360–363, 5.

Rigo Wenning, Matthias Schunter, Lorrie Cranor, B. Dobbs, S. Egelman, G. Hogben, J. Humphrey, M. Langheinrich, M. Marchiori, M. Presler-Marshall, J. Reagle, and D. A. Stampley. 2006. The platform for privacy preferences 1.1 (P3P 1.1) specification.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th World Wide Web Conference*, WWW '13, pages 133–143.

Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium*, USENIX Security '14, pages 1–16. USENIX Association, August.