Unsupervised Alignment of Privacy Policies using Hidden Markov Models

Rohan Ramanath Fei Liu Norman Sadeh Noah A. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{rrohan,feiliu,sadeh,nasmith}@cs.cmu.edu

Abstract

To support empirical study of online privacy policies, as well as tools for users with privacy concerns, we consider the problem of aligning sections of a thousand policy documents, based on the issues they address. We apply an unsupervised HMM; in two new (and reusable) evaluations, we find the approach more effective than clustering and topic models.

1 Introduction

Privacy policy documents are verbose, often esoteric legal documents that many people encounter as clients of companies that provide services on the web. McDonald and Cranor (2008) showed that, if users were to read the privacy policies of every website they access during the course of a year, they would end up spending a substantial amount of their time doing just that and would often still not be able to answer basic questions about what these policies really say. Unsurprisingly, many people do not read them (Federal Trade Commission, 2012).

Such policies therefore offer an excellent opportunity for NLP tools that summarize or extract key information that (i) helps users understand the implications of agreeing to these policies and (ii) helps legal analysts understand the contents of these policies and make recommendations on how they can be improved or made more clear. Past applications of NLP have sought to parse privacy policies into machine-readable representations (Brodie et al., 2006) or extract subpolicies from larger documents (Xiao et al., 2012). Machine learning has been applied to assess certain attributes of policies (Costante et al., 2012; Ammar et al., 2012; Costante et al., 2013; Zimmeck and Bellovin, 2013).

This paper instead analyzes policies in aggregate, seeking to *align* sections of policies. This task is motivated by an expectation that many policies will address similar issues,[1] such as collection of a user's contact, location, health, and financial information, sharing with third parties, and deletion of data. This expectation is supported by recommendation by privacy experts (Gellman, 2014) and policymakers (Federal Trade Commission, 2012); in the financial services sector, the Gramm-Leach-Bliley Act *requires* these institutions to address a specific set of issues. Aligning policy sections is a first step toward our aforementioned summarization and extraction goals.

We present the following contributions:

- A new corpus of over 1,000 privacy policies gathered from widely used websites, manually segmented into subtitled sections by crowdworkers (§2).
- An unsupervised approach to aligning the policy sections based on the issues they discuss. For example, sections that discuss "user data on the company's server" should be grouped together. The approach is inspired by the application of hidden Markov models to sequence alignment in computational biology (Durbin et al., 1998; §3).
- Two reusable evaluation benchmarks for the resulting alignment of policy sections (§4). We demonstrate that our approach outperforms naïve methods (§5).

Our corpus and benchmarks are available at http://usableprivacy.org/data.

2 Data Collection

We collected 1,010 unique privacy policy documents from the top websites ranked by Alexa.com.[2] These policies were collected during a period of six weeks during December 2013 and January 2014. They are a snapshot of privacy policies of mainstream websites covering fifteen

[1] Personal communication, Joel Reidenberg.
[2] http://www.alexa.com

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 605–610,
Baltimore, Maryland, USA, June 23-25 2014.

Business	Computers	Games	Health
Home	News	Recreation	Shopping
Arts	Kids and Teens	Reference	Regional
Science	Society	Sports	

Table 1: Fifteen website categories provided by Alexa.com. We collect privacy policies from the top 100 websites in each.

of Alexa.com's seventeen categories (Table 1).[3]

Finding a website's policy is not trivial. Though many well-regulated commercial websites provide a "privacy" link on their homepages, not all do. We found university websites to be exceptionally unlikely to provide such a link. Even once the policy's URL is identified, extracting the text presents the usual challenges associated with scraping documents from the web. Since every site is different in its placement of the document (e.g., buried deep within the website, distributed across several pages, or mingled together with Terms of Service) and format (e.g., HTML, PDF, etc.), and since we wish to preserve as much document structure as possible (e.g., section labels), full automation was not a viable solution.

We therefore crowdsourced the privacy policy document collection using Amazon Mechanical Turk. For each website, we created a HIT in which a worker was asked to copy and paste the following privacy policy-related information into text boxes: (i) privacy policy URL; (ii) last updated date (or effective date) of the current privacy policy; (iii) privacy policy full text; and (iv) the section subtitles in the top-most layer of the privacy policy. To identify the privacy policy URL, workers were encouraged to go to the website and search for the privacy link. Alternatively, they could form a search query using the website name and "privacy policy" (e.g., "Amazon.com privacy policy") and search in the returned results for the most appropriate privacy policy URL. Given the privacy policy full text and the section subtitles, we partition the full privacy document into different sections, delimited by the section subtitles. A privacy policy is then converted into XML.

Each HIT was completed by three workers, paid \$0.05, for a total cost of \$380 (including Amazon's surcharge).

[3]The "Adult" category was excluded; the "World" category was excluded since it contains mainly popular websites in different languages, and we opted to focus on policies in English in this first stage of research, though mulitlingual policy analysis presents interesting challenges for future work.

3 Approach

Given the corpus of privacy policies described in §2, we designed a model to efficiently infer an alignment of policy sections. While we expect that different kinds of websites will likely address different privacy issues, we believe that many policies will discuss roughly the same set of issues. Aligning the policies is a first step in a larger effort to (i) automatically analyze policies to make them less opaque to users and (ii) support legal experts who wish to characterize the state of privacy online and make recommendations (Costante et al., 2012; Ammar et al., 2012; Costante et al., 2013).

We are inspired by multiple sequence alignment methods in computational biology (Durbin et al., 1998) and by Barzilay and Lee (2004), who described a hidden Markov model (HMM) for document content where each state corresponds to a distinct topic and generates sentences relevant to that topic according to a language model. We estimate an HMM-like model on our corpus, exploiting similarity across privacy policies to the extent it is evident in the data. In our formulation, each hidden state corresponds to an issue or topic, characterized by a distribution over words and bigrams appearing in privacy policy sections addressing that issue. The transition distribution captures tendencies of privacy policy authors to organize these sections in similar orders, though with some variation.

The generative story for our model is as follows. Let $\mathcal{S}$ denote the set of hidden states.

1. Choose a start state y_1 from $\mathcal{S}$ according to the start-state distribution.
2. For $t = 1, 2, \ldots$, until y_t is the stopping state:

(a) Sample the tth section of the document by drawing a bag of terms, $\boldsymbol{o}_t$, according to the emission multinomial distribution for state y_t. Note the difference from traditional HMMs, in which a *single* observation symbol is drawn at each time step. $\boldsymbol{o}_t$ is generated by repeatedly sampling from a distribution over terms that includes all unigrams and bigrams except those that occur in fewer than 5% of the documents and in more than 98% of the documents. This filtering rule was designed to eliminate uninformative stopwords as well as company-specific terms (e.g., the name of the company).[4]

[4]The emission distributions are not a proper language

Category	Websites with privacy URL	Unique privacy policies	Unique privacy policies w/ date	Ave. sections per policy	Ave. tokens per policy
Arts	94	80	72	11.1 (± 3.8)	2894 (± 1815)
Business	100	95	75	10.1 (± 4.9)	2531 (± 1562)
Computers	100	78	62	10.7 (± 4.9)	2535 (± 1763)
Games	92	80	51	10.2 (± 4.9)	2662 (± 2267)
Health	92	86	57	10.0 (± 4.4)	2325 (± 1891)
Home	100	84	68	11.5 (± 3.8)	2493 (± 1405)
Kids and Teens	96	86	62	10.3 (± 4.5)	2683 (± 1979)
News	96	91	68	10.7 (± 3.9)	2588 (± 2493)
Recreation	98	97	67	11.9 (± 4.5)	2678 (± 1421)
Reference	84	86	55	9.9 (± 4.1)	2002 (± 1454)
Regional	98	91	72	11.2 (± 4.2)	2557 (± 1359)
Science	71	75	49	9.2 (± 4.1)	1705 (± 1136)
Shopping	100	99	84	12.0 (± 4.1)	2683 (± 1154)
Society	96	94	65	10.2 (± 4.6)	2505 (± 1587)
Sports	96	62	38	10.9 (± 4.0)	2222 (± 1241)
Average	94.2	85.6	63.0	10.7 (± 4.3)	2471 (± 1635)

Table 2: Statistics of each website category, including (i) the number of websites with an identified privacy policy link; (ii) number of unique privacy policies in each category (note that in rare cases, multiple unique privacy policies were identified for the same website, e.g., a website that contains links to both new and old versions of its privacy policy); (iii) number of websites with an identified privacy modification date; (iv) average number of sections per policy; (v) average number of tokens per policy.

(b) Sample the next state, y_{t+1}, according to the transition distribution over $\mathcal{S}$.

This model can nearly be understood as a hidden *semi*-Markov model (Baum and Petrie, 1966), though we treat the section lengths as observable. Indeed, our model does not even generate these lengths, since doing so would force the states to "explain" the length of each section, not just its content. The likelihood function for the model is shown in Figure 1.

The parameters of the model are almost identical to those of a classic HMM (start state distribution, emission distributions, and transition distributions), except that emissions are characterized by multinomial rather than a categorical distributions. These are learned using Expectation-Maximization, with a forward-backward algorithm to calculate marginals (E-step) and smoothed maximum likelihood estimation for the M-step (Rabiner, 1989). After learning, the most probable assignment of a policy's sections to states can be recovered using a variant of the Viterbi algorithm.

We consider three HMM variants. "Vanilla" allows all transitions. The other two posit an ordering on the states $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, and restrict the set of transitions that are possible, imposing bias on the learner. "All Forward" only allows s_k to transition to $\{s_k, s_{k+1}, \ldots, s_K\}$. "Strict Forward" only allows s_k to transition to s_k or s_{k+1}.

models (e.g., a bigram may be generated by as many as three draws from the emission distribution: once for each unigram it contains and once for the bigram).

4 Evaluation

Developing a gold-standard alignment of privacy policies would either require an interface that allows each annotator to interact with the entire corpus of previously aligned documents while reading the one she is annotating, or the definition (and likely iterative refinement) of a set of categories for manually labeling policy sections. These were too costly for us to consider, so we instead propose two generic methods to evaluate models for sequence alignment of a collection of documents with generally similar content. Though our model (particularly the restricted variants) treats the problem as one of *alignment*, our evaluations consider *groupings* of policy sections. In the sequel, a grouping on a set X is defined as a collection of subsets $X_i \subseteq X$; these may overlap (i.e., there might be $x \in X_i \cap X_j$) and need not be exhaustive (i.e., there might be $x \in X \setminus \bigcup_i X_i$).

4.1 Evaluation by Human QA

This study was carried out as part of a larger collaboration with legal scholars who study privacy. In that work, we have formulated a set of nine multiple choice questions about a single policy that ask about collection of contact, location, health, and financial information, sharing of each with

$$P_{\boldsymbol{\pi},\boldsymbol{\eta},\boldsymbol{\gamma}}\left(\langle y_t, \boldsymbol{o}_t\rangle_{t=1}^{n} \mid \langle \ell_t\rangle_{t=1}^{n}\right) = \pi(y_1)\prod_{t=1}^{n}\left(\prod_{i=1}^{\ell_t}\eta(o_{t,i} \mid y_i)\right)\gamma(y_{t+1} \mid y_t)$$

Figure 1: The likelihood function for the alignment model (one privacy policy). y_t is the hidden state for the tth section, $\boldsymbol{o}_t$ is the bag of unigram and bigram terms observed in that section, and ℓ_t is the size of the bag. Start-state, emission, and transition distributions are denoted respectively by π, η, and γ. y_{n+1} is the silent stopping state.

third parties, and deletion of data.[5] The questions were inspired primarily by the substantive interest of these domain experts—not by this particular algorithmic study.

For thirty policies, we obtained answers from each of six domain experts who were not involved in designing the questions. For the purposes of this study, the experts' answers are not important. In addition to answering each question for each policy, we also asked each expert to copy and paste the text of the policy that contains the answer. Experts were allowed to select as many sections for each question as they saw fit, since answering some questions may require synthesizing information from different sections.

For each of the nine questions, we take the union of all policy sections that contain text selected by any annotator as support for her answer. This results in nine groups of policy sections, which we call **answer-sets** denoted $A_1, \ldots, A_9$. Our method allows these to overlap (63% of the sections in any A_i occurred in more than one A_i), and they are not exhaustive (since many sections of the policies were not deemed to contain answers to any of the nine questions by any expert).

Together, these can be used as a gold standard grouping of policy sections, against which we can compare our system's output. To do this, we define the set of section *pairs* that are grouped together in answer sets, $G = |\{\langle a, b\rangle \mid \exists A_i \ni a, b\}|$, and a similar set of pairs H from a model's grouping. From these sets, we calculate estimates of precision ($|G \cap H|/|H|$) and recall ($|G \cap H|/|G|$).

One shortcoming of this approach, for which the second evaluation seeks to compensate, is that a very small, and likely biased, subset of the policy sections is considered.

4.2 Evaluation by Direct Judgment

We created a separate gold standard of judgments of pairs of privacy policy sections. The data selected for judgment was a sample of pairs stratified by a simple measure of text similarity. We derived unigram tfidf vectors for each section in each of 50 randomly sampled policies per category. We then binned *pairs* of sections by cosine similarity (into four bins bounded by 0.25, 0.5, and 0.75). We sampled 994 section pairs uniformly across the 15 categories' four bins each.

Crowdsourcing was used to determine, for each pair, whether the two sections should be grouped together. A HIT consisted of a pair of policy sections and a multiple choice question, "After reading the two sections given below, would you say that they broadly discuss the same topic?" The possible answers were:

1. Yes, both the sections essentially convey the same message in a privacy policy.
2. Although, the sections do not convey the same message, the broadly discuss the same topic. (For ease of understanding, some examples of content on "the same topic" were included.)
3. No, the sections discuss two different topics.

The first two options were considered a "yes" for the majority voting and for defining a gold standard. Every section-pair was annotated by at least three annotators (as many as 15, increased until an absolute majority was reached). Turkers with an acceptance rate greater than 95% with an experience of at least 100 HITs were allowed and paid $0.03 per annotation. The total cost including some initial trials was $130. 535 out of the 994 pairs were annotated to be similar in topic. An example is shown in Figure 2.

As in §4.1, we calculate precision and recall on pairs. This does not penalize the model for grouping together a "no" pair; we chose it nonetheless because it is interpretable.

5 Experiment

In this section, we evaluate the three HMM variants described in §3, and two baselines, using the methods in §4. All of the methods require the specification of the number of groups or hidden states, which we fix to ten, the average number of sections per policy.

[5]The questions are available in an online appendix at `http://usableprivacy.org/data`.

Section 5 of `classmates.com`*:*

[46 words] . . . You may also be required to use a password to access certain pages on the Services where certain types of your personal information can be changed or deleted. . . . [113 words]

Section 2 of `192.com`*:*

[50 words] . . . This Policy sets out the means by which You can have Your Personal Information removed from the Service. 192.com is also committed to keeping Personal Information of users of the Service secure and only to use it for the purposes set out in this Policy and as agreed by You. . . . [24 words]

Figure 2: Selections from sections that discuss the issue of "deletion of personal information" and were labeled as discussing the same issue by crowdworkers. Both naïve grouping and LDA put them in two different groups, but the Strict Forward variant of our model correctly groups them together.

	Precision		Recall		F_1	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Clust.	0.63	–	0.30	–	0.40	–
LDA	0.56	0.03	0.20	0.05	0.29	0.06
Vanilla	0.62	0.04	0.41	0.04	0.49	0.03
All F.	0.63	0.03	0.47	0.12	0.53	0.06
Strict F.	0.62	0.05	0.46	0.18	0.51	0.07
Clust.	0.62	–	0.23	–	0.34	–
LDA	0.57	0.03	0.18	0.01	0.28	0.02
Vanilla	0.57	0.01	0.30	0.03	0.39	0.02
All F.	0.58	0.02	0.32	0.06	0.41	0.04
Strict F.	0.58	0.03	0.32	0.14	0.40	0.08

Table 3: Evaluation by human QA (above) and direct judgment (below), aggregated across ten independent runs where appropriate (see text). Vanilla, All F(orward), and Strict F(orward) are three variants of our HMM.

Baselines. Our first baseline is a greedy divisive clustering algorithm[6] to partition the policy sections into ten clusters. In this method, the desired K-way clustering solution is computed by performing a sequence of bisections. The implementation uses unigram features and cosine similarity. Our second baseline is latent Dirichlet allocation (LDA; Blei et al., 2003), with ten topics and online variational Bayes for inference (Hoffman et al., 2010).[7] To more closely match our models, LDA is given access to the same unigram and bigram tokens.

Results. Table 3 shows the results. For LDA and the HMM variants (which use random initialization), we report mean and standard deviation across ten independent runs. All three variants of the HMM improve over the baselines on both tasks, in terms of F_1. In the human QA evaluation, this is mostly due to recall improvements (i.e., more pairs of sections relevant to the same policy question were grouped together).

The three variants of the model performed similarly on average, though Strict Forward had very high variance. Its maximum performance across ten runs was very high (67% and 53% F_1 on the two tasks), suggesting the potential benefits of good initialization or model selection.

[6] As implemented in CLUTO, `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

[7] As implemented in gensim (Řehůřek and Sojka, 2010).

6 Conclusion

We considered the task of aligning sections of a collection of roughly similarly-structured legal documents, based on the issues they address. We introduced an unsupervised model for this task along with two new (and reusable) evaluations. Our experiments show the approach to be more effective than clustering and topic models. The corpus and evaluation data have been made available at `http://usableprivacy.org/data` . In future work, policy section alignments will be used in automated analysis to extract useful information for users and privacy scholars.

Acknowledgments

The authors gratefully acknowledge helpful comments from Lorrie Cranor, Joel Reidenberg, Florian Schaub, and several anonymous reviewers. This research was supported by NSF grant SaTC-1330596.

References

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A. Smith. 2012. Automatic categorization of privacy policies: A pilot study. Technical Report CMU-LTI-12-019, Carnegie Mellon University.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proc. of HLT-NAACL*.

Leonard E. Baum and Ted Petrie. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Carolyn A. Brodie, Clare-Marie Karat, and John Karat. 2006. An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. In *Proc. of the Symposium on Usable Privacy and Security*.

Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In *Proc. of the ACM Workshop on Privacy in the Electronic Society*.

Elisa Costante, Jerry Hartog, and Milan Petkovi. 2013. What websites know about you. In Roberto 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 Berlin Heidelberg.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

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

Robert Gellman. 2014. Fair information practices: a basic history (v. 2.11). Available at `http://www.bobgellman.com/rg-docs/rg-FIPShistory.pdf`.

Matthew D Hoffman, David M Blei, and Francis R Bach. 2010. Online learning for latent Dirichlet allocation. In *NIPS*.

Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3).

Lawrence Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proc. of the LREC Workshop on New Challenges for NLP Frameworks*.

Xusheng Xiao, Amit Paradkar, Suresh Thummalapenta, and Tao Xie. 2012. Automated extraction of security policies from natural-language software documents. In *Proc. of the ACM SIGSOFT International Symposium on the Foundations of Software Engineering*.

Sebastian Zimmeck and Steven M. Bellovin. 2013. Machine learning for privacy policy.