A Step Towards Usable Privacy Policy: Automatic Alignment of Privacy Statements

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

Abstract

With the rapid development of web-based services, concerns about user privacy have heightened. The privacy policies of online websites, which serve as a legal agreement between service providers and users, are not easy for people to understand and therefore offer an opportunity for natural language processing. In this paper, we consider a corpus of these policies, and tackle the problem of aligning or grouping segments of policies based on the privacy issues they address. A dataset of pairwise judgments from humans is used to evaluate two methods, one based on clustering and another based on a hidden Markov model. Our analysis suggests a five-point gap between system and median-human levels of agreement with a consensus annotation, of which half can be closed with bag of words representations and half requires more sophistication.

1 Introduction

Privacy policies are legal documents, authored by privacy lawyers to protect the interests of companies offering services through the web. According to a study conducted by McDonald and Cranor (2008), if every internet user in the U.S. read the privacy notice of each new website she visited, it would take the nation 54 billion hours annually to read privacy policies. It is not surprising that they often go unread (Federal Trade Commission, 2012).

Users, nonetheless, might do well to understand the implications of agreeing to a privacy policy, and might make different choices if they did. Researchers in the fields of internet privacy and security have made various attempts to standardize the format of privacy notices, so that they are easier to understand and to allow the general public to have better control of their personal information. An early effort is the Platform for Privacy Preferences Project (P3P), which defines a machine-readable language that enables the websites to explicitly declare their intended use of personal information (Cranor, 2002). Many other studies primarily focus on the qualitative perspective of policies and use tens of carefully selected privacy notices. For example, Kelley et al. (2010) proposed a "nutrition label" approach that formalizes the privacy policy into a standardized table format. Breaux et al. (2014) map privacy requirements encoded in text to a formal logic, in order to detect conflicts in requirements and trace data flows (e.g., what data might be collected, to whom the data will be transferred and for what purposes).

The need for automatically or semi-automatically generating simple, easy-to-digest privacy summaries is further exacerbated by the emergence of the mobile Web and the Internet of Things, with early efforts in this area including the use of static analysis to identify sensitive data flows in mobile apps (Lin et al., 2012) and the development of mobile app privacy profiles (Liu et al., 2014).

Increased automation for such efforts motivates our interest in privacy policies as a text genre for NLP, with the general goal of supporting both user-oriented tools that interpret policies and studies of the contents of policies by legal scholars.

In this paper, we start with a corpus of 1,010 policies collected from widely-used websites (Ramanath et al., 2014),[1] and seek to automatically align segments of policies. We believe this is a worthwhile first

[1] http://www.usableprivacy.org/data

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 884–894, Dublin, Ireland, August 23-29 2014.

Amazon.com Privacy Notice	**Walmart Privacy Policy**
... **What About Cookies?** Cookies are unique identifiers that we transfer to your device to enable our systems to recognize your device and to provide features such as 1-Click purchasing, Recommended for You, personalized advertisements on other Web sites... ...Because cookies allow you to take advantage of some of Amazon.com's essential features, we recommend that you leave them turned on. For instance, if you block or otherwise reject our cookies, you will not be able to add items to your Shopping Cart, proceed to Checkout, or use any Amazon.com products and services that require you to Sign in...	... **Information We Collect** ...We use "cookies" to recognize you as you use or return to our sites. This is done so that we can provide a continuous and more personalized shopping experience for you. A cookie is a small text file that a website or email may save to your browser and store on your hard drive... **Your Choices** ...You may exercise choices related to our online operations and advertising. For instance, you can choose to browse our websites without accepting cookies. Please know that cookies allow us to recognize you from page to page, and they support your transactions with us. Without cookies enabled, you will still be able to browse our websites, but will not be able to complete a purchase or take advantage of certain website features...

Table 1: Example privacy statements from Amazon.com (left) and Walmart.com (right). The statements are concerned with the websites' cookie policy. The top-most level section subtitles are shown in bold.

step toward interpretation of the documents of direct interest here, and also that automatic alignment of a large set of similarly-constructed documents might find application elsewhere.

Consider the example in Table 1, where we show privacy statements from Amazon.com[2] and Walmart.com.[3] These statements are concerned with the usage of cookies—small data files transferred by a website to the user's computer hard drive—often used for tracking a user's browsing behavior. Cookies are one issue among many that are addressed by privacy policies; by aligning segments by issue, across policies, we can begin to understand the range of policy approaches for each issue.

We contribute pairwise annotations of segment pairs drawn from different policies, for use in evaluating the quality of alignments, an analysis of the inter-annotator reliability, and an experimental assessment of three alignment methods, one based on clustering and two based on a hidden Markov model. This paper's results refine the findings of Ramanath et al. (2014). Our key finding is that these unsupervised methods reach far better agreement with the consensus of crowdworkers than originally estimated, and that the gap between these methods and the "median" crowdworker is about half due to the greedy nature of such methods and about half due to the bag of words representation.

2 Privacy Dataset and Annotations

For completeness, we review the corpus of privacy policies presented by Ramanath et al. (2014), and then present the new annotations created for evaluation of alignment.

2.1 Corpus

We collected 1,010 unique privacy policy documents from the top websites ranked by Alexa.com.[4] 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 of Alexa.com's seventeen categories (Table 2).[5]

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.

[2] https://www.amazon.com/gp/help/customer/display.html?nodeId=468496

[3] http://corporate.walmart.com/privacy-security/walmart-privacy-policy

[4] http://www.alexa.com

[5] 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 multilingual policy analysis presents interesting challenges for future work.

	Sections		Paragraphs			Sections		Paragraphs	
Category	Count	Length	Count	Length	Category	Count	Length	Count	Length
Arts	11.1	254.8	39.2	72.1	Recreation	11.9	218.8	38.5	67.4
Business	10.0	244.2	37.6	65.1	Reference	9.7	179.4	26.2	66.3
Computers	10.5	213.4	34.4	65.4	Regional	10.2	207.7	36.0	59.1
Games	10.0	244.1	34.9	70.1	Science	8.7	155.0	22.1	61.0
Health	9.9	228.2	32.4	69.4	Shopping	11.9	213.9	39.3	64.8
Home	11.6	201.5	32.4	72.0	Society	9.8	230.8	32.6	69.3
Kids and Teens	9.6	231.5	32.3	68.6	Sports	10.1	217.1	29.1	75.6
News	10.3	248.4	35.5	72.4	**Average**	**10.4**	**221.9**	**34.1**	**68.0**

Table 2: Fifteen website categories, average number of sections and paragraphs per document in that category, and average length in word tokens.

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. Each HIT was completed by three workers, paid $0.05, for a total cost of $380 (including Amazon's surcharge). After excluding dupliates, the dataset contains 1,010 unique documents.[6]

Given the privacy policy full text and the section subtitles, we partition the full privacy document into different sections, delimited by the section subtitles. To generate paragraphs, we break the sections by lines, and consider each line as a paragraph. We require a paragraph to end with a period, if not, it will be concatenated with the next paragraph. Using this partition scheme, sections contain 12 sentences on average; and paragraphs contain 4 sentences on average. More statistics are presented in Table 2.

2.2 Pairwise Annotations

Ramanath et al. (2014) described an evaluation method in which pairs of privacy policy sections were annotated by crowdworkers.[7] A sample of section pairs from different policies was drawn, stratified by cosine similarity of unigram tfidf vectors. In a single task, a crowdworker was asked whether two sections broadly discussed the same topic. The question was presented alongside three answer options, essentially a strong yes, a yes, and a no. In that initial exploration, each item was annotated at least three times, and up to fifteen, until an absolute majority was reached.

The annotations conducted for this study were done somewhat differently. Our motivations were to enable a more careful exploration of inter-annotator agreement, which was complicated in the earlier work by the variable number of annotations per pair, from three to fifteen. We also sought to explore a more fine-grained problem at the paragraph level.

We sampled 1,000 document pairs from each of the 15 categories, then generated pairs (separately of sections and of paragraphs) by choosing one at random from each document. In total, 1,278,204 section pairs and 7,968,487 paragraph pairs were produced. These pairs were stratified by cosine similarity intervals: [0, 0.25], (0.25, 0.5], (0.5, 0.75], (0.75, 1], as in Ramanath et al. (2014). We sampled 250 pairs from each interval, resulting in 1,000 pairs each of sections and paragraphs.

These pairs were annotated on Amazon Mechanical Turk. The crowdworkers were instructed to carefully read the privacy statements and answer a "yes/no" question, indicating whether the two texts are discussing the same privacy issue or not. Several key privacy issues are provided as examples, including

[6] Note that different websites may be covered by the same privacy policy provided by the parent company. For example, `espn.go.com`, `abc.go.com`, and `marvel.com` are all covered under the Walt Disney privacy policy.

[7] Another evaluation, based on text selected by humans in a separate, unrelated task, was also explored. Because such an evaluation seems less broadly applicable, we did not pursue it here.

	Sections					Paragraphs				
Cosine similarity:	[0, .25]	(.25, .5]	(.5, .75]	(.75, 1]	All	[0, .25]	(.25, .5]	(.5, .75]	(.75, 1]	All
5 workers agree	36.4	**12.4**	28.0	85.2	40.5	42.4	**12.0**	32.8	77.6	41.2
4 workers agree	42.8	**42.4**	42.0	13.6	35.2	39.6	**36.8**	35.6	17.6	32.4
3 workers agree	20.8	**45.2**	30.0	1.2	24.3	18.0	**51.2**	31.6	4.8	26.4
Consensus-yes	4.4	**45.2**	87.2	99.2	59.0	9.2	**66.0**	88.8	98.0	65.5
Consensus-no	95.6	**54.8**	12.8	0.8	41.0	90.8	**34.0**	11.2	2.0	34.5

Table 3: Inter-annotator agreement of section and paragraph pairs.

collection of personal information, sharing of information with third parties, cookies and other tracking techniques, data security, children policies, and contact of the websites. To encourage the crowdworkers to carefully read the privacy statements, we also asked them to copy and paste 1–3 keywords from each section/paragraph, before answering the question.[8] Each section/paragraph pair was judged by five crowdworkers and was rewarded \$0.05. In total, \$550 was spent on the annotations.

On average, it took a crowdworker 2.15 minutes to complete a section pair and 1.67 minutes for a paragraph pair. Interestingly, although a section is roughly three times the length of a paragraph (see Table 2), the time spent on annotation is not proportional to the text length.

In Table 3, we present the inter-annotator agreement results for section and paragraph pairs, broken down by cosine-similarity bin and by the majority answer. 75.7% (73.6%) of section (paragraph) pairs were agreed upon by four or more out of five annotators. Unsurprisingly, disagreement is greatest in the $(.25, .5]$ similarity bin. Cosine similarity is a very strong predictor of the consensus answer (Pearson correlation 0.72 for section pairs, 0.67 for paragraphs, on this stratified sample).

Ramanath et al. (2014) considered only sections. A different method was used to obtain consensus annotations; we simply kept adding annotators to a pair until consensus was reached. For a fair comparison with the new data, we calculated pairwise agreement among three annotators per item, randomly selected if there were more than three to choose from. On the old section-level data, this was 60.5%; on the new data, it was 71.3% (using five annotators). Although a controlled experiment in the task setup was not conducted, we take this as a sign that our binary question with keywords led to a higher quality set of annotations than the three-way question in the older data. Our experiments in this paper use only the new data.

2.3 Discussion

We had expected higher agreement at the paragraph level, since paragraphs are shorter, presumably easier to read and compare, and presumably more focused on a smaller number of issues. This was not borne out empirically, though a slightly different analysis presented in §4.2 suggests that, among crowdworkers who completed ten or more tasks, paragraphs were easier to agree on.

Privacy policies are generally written by attorneys with expertise in privacy law, though there are automatic generation solutions available that allow a non-expert to quickly fill in a template to create a policy document.[9] Example 1 in Table 4 shows a case of very high text overlap (five out of five annotators agreed on a "yes" answer for this pair). While this kind of localized alignment is not our aim here, we believe that such "boilerplate" text, to the extent that it occurs in large numbers of policies, will make automatic alignment easier.

A case where annotators seem not to have understood, or not taken care to read carefully, is illustrated by Example 2 in Table 4. Both sections describe "opt-out" options for unsubscribing from mailing lists that send promotional messages, though the first is more generally about "communications" and the second only addresses email. Three out of five crowdworkers labeled this example with"no." Achieving better consensus might require more careful training of annotators about a predefined set of concepts at the right granularity.

[8] We have not used these keywords for any other purpose.

[9] For example: `http://www.rendervisionsconsulting.com/blog/wp-content/uploads/2011/09/Privacy-policy-solutions-list_rvc.pdf`

Example 1	Example 2
Policy excerpt from Urban Outfitters website: To serve you better, we may combine information you give us online, in our stores or through our catalogs. We may also combine that information with publicly available information and information we receive from or cross-reference with our Select Partners and others. We use that combined information to enhance and personalize the shopping experience of you and others with us, to communicate with you about our products and events that may be of interest to you, and for other promotional purposes. **Policy excerpt from Williams-Sonoma website:** To serve you better and improve our performance, we may combine information you give us online, in our stores or through our catalogs. We may also combine that information with publicly available information and information we receive from or cross-reference with select partners and others. By combining this information we are better able to communicate with you about our products, special events and promotional purposes and to personalize your shopping experience.	**Policy excerpt from IKEA website:** What if I prefer not to receive communications from IKEA? If you prefer not to receive product information or promotions from us by U.S. Mail, please click here. To unsubscribe from our email list, please follow the opt-out instructions at the bottom of the email you received, or click here and update your profile by deselecting "Please send me: Inspirational emails and updates." **Policy excerpt from Neiman Marcus website:** Emails. You will receive promotional emails from us only if you have asked to receive them. If you do not want to receive email from Neiman Marcus or its affiliates you can click on the "Manage Your Email Preferences" link at the bottom of any email communication sent by us. Choose "Unsubscribe" at the bottom of the page that opens. Please allow us 3 business days from when the request was received to complete the removal, as some of our promotion s may already have been in process before you submitted your request.

Table 4: Privacy policy excerpts. Example 1 (a pair of paragraphs) illustrates the likely use of boilerplate; identical text is marked in gray. Example 2 shows a pair of sections where our intuitions disagree with the annotations.

3 Problem Formulation

Given a collection of privacy policy documents and assuming each document consists of a sequence of naturally-occurring text segments (e.g., sections or paragraphs), our goal is to automatically group the text segments that address the same privacy issue, without pre-specifying the set of such issues. We believe this exemplifies many scenarios where a collection of documents follow a similar content paradigm, such as legal documents and, in some cases, scientific literature. Our interest in algorithms that characterize each individual document's parts in the context of the corpus is inspired by biological sequence alignment in computational biology (Durbin et al., 1998).

In our experiments, we consider a hidden Markov model (HMM) that captures local transitions between topics. The motivation for the HMM is that privacy policies might tend to order issues similarly, e.g., the discussion on "sharing information to third parties" appears to often follow the discussion of "personal information collection." If each of these corresponds to an HMM state, then the regularity in ordering is captured by the transition distribution, and each state is characterized by its emission distribution over words. In this section, we discuss the HMM and two estimation procedures based on Expectation-Maximization (EM) and variational Bayesian (VB) inference.

3.1 Hidden Markov Model

Assume we have a sequence of observed text segments[10] $O = [O_1, O_2, ..., O_T]$, and each O_t represents a text segment in a privacy document ($t \in \{1, 2, ..., T\}$). We denote $O_t = [O_t^1, O_t^2, ..., O_t^{N_t}]$, where each O_t^j corresponds to a word token in the tth text segment; N_t is the total number of word tokens in the segment; T represents the total number of segments in the observation sequence. Each text segment O_t is associated with a hidden state S_t ($S_t \in \{1, 2, \ldots, K\}$, where K is the total number of states). Given an observation sequence O, our goal is to decode the corresponding hidden state sequence S.

We employ a first-order hidden Markov model where the next state depends only on the previous state. A notable difference from the familiar HMM used in NLP (e.g., as used for part-of-speech tagging) is that we allow multiple observation symbols to be emitted from each hidden state. Each symbol corresponds to a word token in the text segment. Hence the likelihood for a single document can be written as:

$$L(\theta, \phi) = \sum_{S \in \{1,...,K\}^T} p(O, S \mid \theta, \phi) = \sum_{S \in \{1,...,K\}^T} \prod_{t=1}^{T+1} \theta_{S_t|S_{t-1}} \prod_{j=1}^{N_t} \phi_{O_t^j|S_t} \qquad (1)$$

[10] We use *segments* to refer abstractly to either sections or paragraphs. In any given instantiation, one or the other is used, never a blend.

E-step:

$$\textit{Forward pass:}\qquad \alpha_1(\cdot) = 1; \quad \alpha_t(k) = \sum\nolimits_{k'=1}^{K} \alpha_{t-1}(k') \cdot \theta_{k|k'} \cdot \prod\nolimits_{j=1}^{N_t} \phi_{O_t^j|k}, \;\; \forall t \in \{2, \ldots, T\}, \forall k \in \{1, \ldots, K\} \quad (2)$$

$$\textit{Backward pass:}\qquad \beta_{T+1}(\cdot) = 1; \quad \beta_t(k) = \sum\nolimits_{k'=1}^{K} \theta_{k'|k} \cdot \prod\nolimits_{j=1}^{N_t} \phi_{O_t^j|k'} \cdot \beta_{t+1}(k'), \;\; \forall t \in \{T, \ldots, 1\}, \forall k \in \{1, \ldots, K\} \quad (3)$$

$$\textit{Likelihood:}\qquad p(O \mid \theta, \phi) = p(O_1, O_2, ..., O_T \mid \theta, \phi) = \sum\nolimits_{k=1}^{K} \alpha_t(k) \cdot \beta_t(k) \; \textit{(for any t)} \quad (4)$$

$$\textit{Posteriors:}\qquad \gamma_t(k) = p(S_t = k \mid O, \theta, \phi) = \frac{\alpha_t(k) \cdot \beta_t(k)}{p(O \mid \theta, \phi)} \quad (5)$$

$$\textit{Pair posteriors:}\qquad \xi_t(k, k') = p(S_t = k, S_{t+1} = k' \mid O, \theta, \phi) = \frac{\alpha_t(k) \cdot \theta_{k'|k} \cdot \left(\prod_{j=1}^{N_{t+1}} \phi_{O_{t+1}^j|k'}\right) \cdot \beta_{t+1}(k')}{p(O \mid \theta, \phi)} \quad (6)$$

M-step (in EM):

$$\textit{Transitions:}\quad \theta_{k'|k} = \frac{\sum_{t=1}^{T} \xi_t(k, k')}{\sum_{t=1}^{T} \sum_{k''=1}^{K} \xi_t(k, k'')}; \qquad \textit{Emissions:}\quad \phi_{v|k} = \frac{\sum_{t=1}^{T} \gamma_t(k) \cdot \sum_{j=1}^{N_t} \mathbf{1}\{O_t^j = v\}}{\sum_{t=1}^{T} \gamma_t(k) \cdot N_t} \quad (7)$$

Variational update (in VB):

$$\theta_{k'|k} = \frac{\exp \Psi \left(\sum_{t=1}^{T} \xi_t(k, k') + \lambda\right)}{\exp \Psi \left(\sum_{t=1}^{T} \sum_{k''=1}^{K} \xi_t(k, k'') + \lambda \cdot K\right)}; \qquad \phi_{v|k} = \frac{\exp \Psi \left(\sum_{t=1}^{T} \gamma_t(k) \cdot \sum_{j=1}^{N_t} \mathbf{1}\{O_t^j = v\} + \lambda'\right)}{\exp \Psi \left(\sum_{t=1}^{T} \gamma_t(k) \cdot N_t + \lambda' \cdot V\right)} \quad (8)$$

Table 5: Equations for parameter estimation of the HMM with multiple emissions at each state and a single sequence. K is the number of states, V is the emission vocabulary size, and T is the length of the sequence in sections. $\Psi(\cdot)$ is the digamma function.

$\theta_{k'|k}$ denotes the probability of transitioning to state k' given that the preceding state is k. $\phi_{v|k}$ denotes the probability that a particular symbol emitted during a visit to state k is the word v. As in standard treatments, we assume an extra final state at the end of the sequence that emits a stop symbol.

Ramanath et al. (2014) considered three variants of the HMM, with different constraints on the transitions, such as a "strict forward" variant that orders the states and only allows transition to "later" states than the current one. In the evaluation against direct human judgments, they found a slight benefit from such constraints, but they increased performance variance considerably. Here we only consider an unconstrained HMM.

3.2 EM and VB

We consider two estimation methods, neither novel. Both are greedy hillclimbing methods that locally optimize functions based on likelihood under the HMM.

The first method is EM, adapted for the multiple emission case; the equations for the E-step (forward-backward algorithm and subsequent posterior calculations) and the M-step are shown in Table 5.

We also consider Bayesian inference, which seeks to marginalize out the parameter values, since we are really only interested in the assignment of sections to hidden states. Further, Bayesian inference has been found favorable on small datasets (Gao and Johnson, 2008). We assume symmetric Dirichlet priors on the transition and emission distributions, parameterized respectively by $\lambda = 1$ and $\lambda' = 0.1$. We apply mean-field variational approximate inference as described by Beal (2003), which amounts to an EM-like procedure. The E-step is identical to EM, and the M-step involves a transformation of the expected counts, shown in Table 5. (We also explored Gibbs sampling; performance was less stable but generally similar; for clarity we do not report the results here.)

3.3 Implementation Details

In modeling, the vocabulary excludes 429 stopwords,[11] words whose document frequency is less than ten, and a set of terms specific to website privacy polices: *privacy*, *policy*, *personal*, *information*, *service*, *web*, *site*, *website*, *com*, and *please*. After lemmatizing, the vocabulary contains V = 2,876 words. We further exclude sections and paragraphs that contain less than 10 words. Many of these are not meaningful statements, e.g., "return to top." This results in 9,935 sections and 27,594 paragraphs in the experiments.

During estimation, we concatenate all segments into a single sequence, delimited by a special boundary symbol. This does not affect the outcome (due to the first-order conditions; it essentially conflates "start" and "stop" states), but gave some efficiency gains in our implementation.

EM or VB iterations continue until one of two stopping criteria is met: either 100 iterations have passed, or the relative change in log-likelihood (or the variational bound in the case of VB) falls below 10^{-4}; this consistently happens within forty iterations.

After estimating parameters, we decode using the Viterbi algorithm.

4 Experiments

Our experiments compare three methods for aligning segments of privacy policies, at both the paragraph and the section level:

- A greedy divising clustering algorithm, as implemented in CLUTO.[12] The algorithm performs a sequence of bisections until the desired number of clusters is reached. In each step, a cluster is selected and partitioned so as to optimize the clustering criterion. CLUTO demonstrated robust performance in several related NLP tasks (Zhong and Ghosh, 2005; Lin and Bilmes, 2011; Chen et al., 2011).

- The Viterbi state assignment from the EM-estimated HMM. We report averaged results over ten runs, with random initialization.

- The Viterbi state assignment after VB inference, using the mean field parameters. We report averaged results over ten runs, with random initialization.

Our evaluation metrics are precision, recall, and F-score on the identification of section or paragraph pairs annotated "yes."

4.1 Results

In Figure 1, we present performance of different algorithms using a range of hidden state values $K \in \{1, 2, \ldots, 20\}$. The top row shows precision, recall and F-scores on section pairs, the bottom row on paragraph pairs.

The algorithms mostly perform similarly. At the section level, we find the clustering algorithm to perform better in terms of F-score than the HMM with larger K; at $K = 10$ the two are very close.[13] CLUTO's best performance, 85%, was achieved by $K = 14$.

At the paragraph level, the HMMs outperform clustering in the $K \in [5, 15)$ range, and this is where the peak F-score is obtained (87%). We do not believe these differences among algorithms are especially important, noting only that the HMM's advantage is that it does not require pairwise similarity calculations between all section pairs.

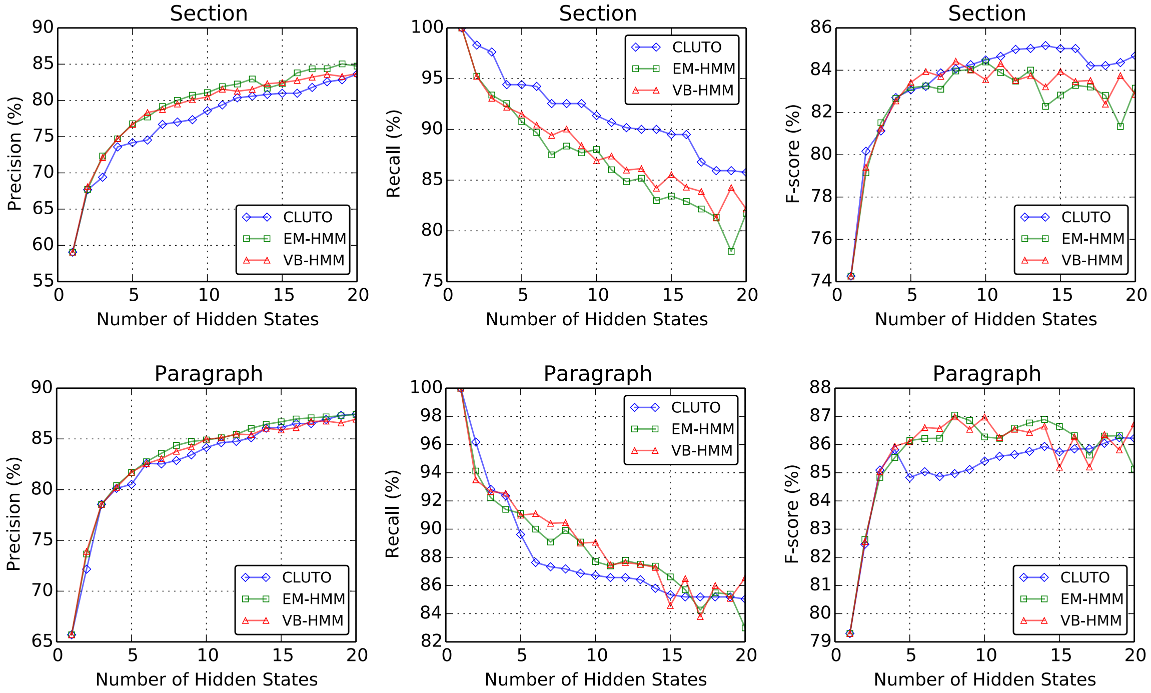

Figure 1: Performance results against pairwise annotations when using different number of hidden states $K \in \{1, \ldots, 20\}$. The top row is at the section level, the bottom row at the paragraph level.

4.2 Upper Bounds

How do these automatic alignment methods compare with the levels of agreement reached among crowdworkers? We consider the agreement rate of each method, at varying values of K, with the majority vote of the annotators. Note that this is distinct from the positive-match–focused precision, recall, and F-score measures presented in §4.1. For each crowdworker who completed ten tasks or more, and therefore for whom we have hope of a reliable estimate, we calculated her agreement rate with the majority. For sections, this set included 65 out of 162 crowdworkers; for paragraphs, 76 out of 197.

In Figure 2 we show the three quartile points for this agreement measure, across the pool of ten-or-more-item crowdworkers, in comparison to the various automatic methods. For sections, our systems perform on par with the 25% of crowdworkers just below the median. For paragraphs, which show a generally higher level of agreement among this subset of crowdworkers, our systems are on par with the lowest 25% of workers. We take all of this to suggest that there is room for improvement in methods overall.

Given the observation in §2 that cosine similarity of two segments' tfidf vectors is a very strong predictor of human agreement on whether they are about the same issue, we also consider a threshold on cosine similarity for deciding whether a pair is about the same issue. This is not a complete solution to the problem of alignment, since pairwise scores only provide groupings if they are coupled with a transitivity constraint. The clustering and HMM methods can be understood as greedy approximations to such an approach. We therefore view cosine similarity thresholding as an upper bound for bag of words representations on the pairwise evaluation task. Figure 2 includes agreement levels for oracle cosine similarity thresholding.[14]

[11] `http://www.lextek.com/manuals/onix/stopwords1.html`

[12] `http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview`

[13] Ramanath et al. (2014) only considered $K = 10$ and found a $K = 10$ HMM to outperform clustering at the section level; the scores reported there, on the earlier dataset, are much lower and not comparable to those reported here. There are numerous differences between the setup here and the earlier one. The most important, we believe, are the improved quality of the dataset and greater care given to preprocessing, most notably the pruning of documents and vocabulary, in the present experiments.

[14] For comparison with the results in §4.1, we found that, for sections, oracle thresholding (at 0.3) achieved F-score of 0.87, and for paragraphs, oracle thresholding (at 0.2) achieved 0.90.

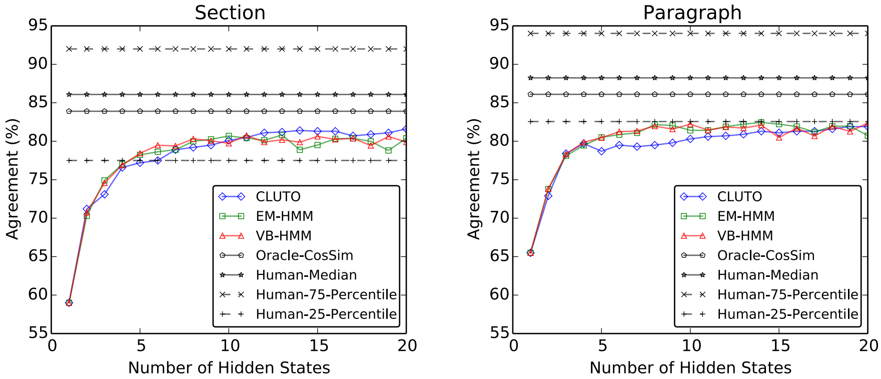

Figure 2: Agreement rates, as compared to crowdworkers and a cosine similarity oracle.

Taken together, this analysis suggests that—in principle—an automated approach based on word-level similarity could close about half of the gap between our methods and median crowdworkers, and further gains would require more sophisticated representations or similarity measures.

5 Related Work

There has been little work on applying NLP to privacy policies. Some have sought to parse privacy policies into machine-readable representations (Brodie et al., 2006) or extract sub-policies from larger documents (Xiao et al., 2012). Machine learning has been applied to assess certain attributes of policies (Costante et al., 2012; Costante et al., 2013), e.g., compliance of privacy policies to legal regulations (Krachina et al., 2007) or simple categorical questions about privacy policies (Ammar et al., 2012; Zimmeck and Bellovin, 2014).

Our alignment-style analysis is motivated by an expectation that many policies will address similar issues,[15] 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. Sadeh et al. (2013) describe our larger research initiative to incorporate automation into privacy policy analysis.

Methodologically, the HMM used above is very similar to extensive previous uses of HMMs for POS tagging (Merialdo, 1994), including with Bayesian inference (Goldwater and Griffiths, 2007; Johnson, 2007; Gao and Johnson, 2008). Bayesian topic models (Blei et al., 2003) are a related set of techniques, and future exploration might consider their use in automatically discovering document sections (Eisenstein and Barzilay, 2008), rather than fixing section or paragraph boundaries.

6 Conclusion

This paper presents an exploration of alignment-by-paragraph and -section of website privacy policies. We contribute an improved annotated dataset for pairwise evaluation of automatic methods and an exploration of clustering and HMM-based alignment methods. Our results show that these algorithms achieve agreement on par with the lower half of crowdworkers, with about half of the difference from the median due to the bag of words representation and half due to the inherent greediness of the methods.

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.

[15] Personal communication, Joel Reidenberg.

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.

Matthew J. Beal. 2003. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Gatsby Computational Neuroscience unit, University College London.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*.

Travis D. Breaux, Hanan Hibshi, and Ashwini Rao. 2014. Eddy, A formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. *Requirements Engineering Journal*.

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 *Proceedings of the Second Symposium on Usable Privacy and Security (SOUPS)*.

Harr Chen, Edward Benson, Tahira Naseem, and Regina Barzilay. 2011. In-domain relation discovery with meta-constraints via posterior regularization. In *Proceedings of ACL-HLT*.

Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness. In *Proceedings 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.

Lorrie Faith Cranor. 2002. *Web Privacy with P3P*. O'Reilly & Associates.

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

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of ACL*.

Federal Trade Commission. 2012. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers. Available at `http://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers`.

Jianfeng Gao and Mark Johnson. 2008. A comparison of Bayesian estimators for unsupervised Hidden Markov model POS taggers. In *Proceedings of EMNLP*.

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

Sharon Goldwater and Thomas L. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*.

Mark Johnson. 2007. Why doesnt EM find good HMM POS-taggers? In *Proceedings of EMNLP–CoNLL*.

Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing privacy notices: An online study of the nutrition label approach. In *Proceedings of CHI*.

Olga Krachina, Victor Raskin, and Katrina Triezenberg. 2007. Reconciling privacy policies and regulations: Ontological semantics perspective. In Michael J. Smith and Gavriel Salvendy, editors, *Human Interface and the Management of Information. Interacting in Information Environments*, pages 730–739. Springer.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of ACL*.

Jialiu Lin, Shahriyar Amini, Jason I. Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*.

Bin Liu, Jialiu Lin, and Norman Sadeh. 2014. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In *Proceedings of WWW*.

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*.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A. Smith. 2014. Unsupervised alignment of privacy policies using hidden Markov models. In *Proceedings of ACL*.

Norman Sadeh, Alessandro Acquisti, Travis Breaux, Lorrie Cranor, Aleecia McDonald, Joel Reidenberg, Noah Smith, Fei Liu, Cameron Russel, 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.

Xusheng Xiao, Amit Paradkar, Suresh Thummalapenta, and Tao Xie. 2012. Automated extraction of security policies from natural-language software documents. In *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering*.

Shi Zhong and Joydeep Ghosh. 2005. Generative model-based document clustering: a comparative study. *Knowledge and Information Systems*, 8(3):374–384.

Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *Proceedings of the 23rd USENIX Security Symposium*.