Towards an Information Type Lexicon for Privacy Policies

Jaspreet Bhatia, Travis D. Breaux
Institute for Software Research
Carnegie Mellon University
Pittsburgh, Pennsylvania, United States
{jbhatia, breaux}@cs.cmu.edu

***Abstract*—Privacy policies serve to inform consumers about a company's data practices, and to protect the company from legal risk due to undisclosed uses of consumer data. In addition, US and EU regulators require companies to accurately describe their practices in these policies, and some laws prescribe how companies should write these policies. Despite these aims, privacy policies are frequently criticized for being vague and uninformative. To support and improve the analysis of privacy policies, we report results from constructing an information type lexicon from manual, human annotations and an entity extractor based on part-of-speech tagging. The lexicon was constructed from 3,850 annotations obtained from crowd workers analyzing 15 privacy policies. An entity extractor was designed to extract entities from these annotations. The extractor succeeds at finding entities in 92% of annotations and the lexicon consists of 725 unique entities. Finally, we measured the terminological reuse across all 15 policies and observed the lexicon has a 31-78% chance of containing a word from any previously seen policy.**

***Index Terms*—requirements extraction, crowdsourcing, natural language processing, privacy**

I. Introduction

Privacy policies describe the information practices for a company, website or one or more products or services. These policies are typically written by legally trained personnel and often aim to answer important questions, such as: what personal information is collected, for what purposes is the information used, and with whom is the information shared? In addition, privacy policies may describe under what conditions an individual may consent to, opt-out of, or opt-in to various data practices. Evidence shows that many people do not read privacy policies [13], and that when they do try, policies are written at a college reading-level [14]. Despite this evidence, privacy policies still play a critical role in data transparency and they are increasingly a source of discussion for companies, regulators and researchers.

Legislators write privacy laws that affect the organization and content of privacy policies. For example, the Health Insurance Portability and Accountability Act (HIPAA) requires healthcare privacy policies to describe patient rights, the Gramm-Leach-Blilely Act (GLBA) requires financial institutions to include example data practices, and the Children's Online Privacy Protection Act (COPPA) requires policies to describe how parental consent can be obtained to protect children under 13 years. Laws from other nations and provinces further impose requirements on policy content.

The manner in which privacy policies describe personal information, both the category of information and the level of detail, can affect the outcome of privacy policy analyses. To support and improve analysis of data practices in privacy policies, we propose to extract "information types," which are noun phrases that describe personal information. We conducted a study to develop an information type lexicon based on privacy policy annotations obtained from crowdsourcing and an entity extractor based on part-of-speech (POS) tagging. With a lexicon of information types, we aim to support richer analysis of policies, such as detecting sector-specific practices concerning sensitive information types, or measuring the degree of ambiguity due to less or more precise use of terms (e.g., contact information versus e-mail address and phone number, which are more specific).

The remainder of this paper is organized as follows: in Section II, we review related work; in Section III, we introduce our crowdsourcing task, the entity extraction method, and the lexicon construction method; in Section IV, we report results; and in Section V, we present discussion and future work.

II. Related Work

We now review related work in requirements extraction from text, use of natural language processing (NLP) and machine learning (ML) for requirements analysis and from crowdsourcing annotations.

A. Requirements Extraction

The translation from text to formal and semi-formal specifications has long been a challenge. Abbot first examined mining program descriptions from text for object-oriented design [1]. Later, Antón introduced the GBRAM and heuristics to extract goal specifications from text. Goals range from high- and low-level actions to be maintained, achieved and avoided by the system [9]. Antón and Earp applied GBRAM to mine privacy goals from privacy policies [2], and Breaux et al. later showed how to extract data flow requirements from privacy policies [6]. In this paper, we extend this prior work with a narrow focus on extracting information types from privacy policy statements and building an information type lexicon.

B. Natural Language Processing and Machine Learning for Requirements Analysis.

According to Berry [4], majority of the requirements are written in natural language, hence it becomes important to

Accepted for publication by IEEE. © 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

develop techniques that automatically analyze requirements. Massey et al. use text-mining methods to identify the presence of software requirement artifacts in policy documents [12]. The Privee architecture is based on crowdsourcing and automatic classification of privacy policies based on the analysis of "essential" policy terms [17]. Breaux et al. describe the Eddy language for modeling privacy policies, which include information types as principal Eddy expressions [7]. We believe the work by Massey et al. and the developers of Privee can benefit from an information type lexicon that maps terms across policies to single entities.

In contrast, the policy workbench SPARCLE developed at IBM Watson, takes as input the natural language policy document and uses NLP techniques to parse the policy text, identify policy elements and generates machine-readable XML version of the policy [8]. SPARCLE aims to simplify policies by making them machine-readable. We expect that our lexicon could be used to help policy authors choose terms to which their data practices apply, including those practices expressed in SPARCLE.

C. Crowdsourcing Annotations and Extraction

Crowdsourcing facilitates tackling problems that remain hard to solve with automated methods by leveraging human intelligence, typically provided by non-experts [15]. Crowdsourced annotations from non-experts have also been shown to be comparable to expert annotations for certain annotation tasks, such as word similarity, word sense disambiguation and textual entailment recognition [16]. Crowdsourcing has also been employed for requirements elicitation: StakeRare uses social networks and collaborative filtering to elicit and prioritize user requirements [11]. Breaux & Schaub used crowdsourcing to extract privacy requirements from websites' privacy policies largely matching expert performance but resulting in larger coverage [5].

In order to leverage the potential of crowdsourcing for annotating and extracting natural language text a number of challenges need to be addressed. André et al. [3] note the major challenges as having non-experts perform the annotations, a transient workforce, and the need to resolve conflicting and potentially erroneous annotations.

III. Lexicon Construction Framework

We now describe the lexicon construction framework that consists of three parts: a crowd worker task to obtain manual annotations, the entity extraction method, and the lexicon construction method. Figure 1 shows our framework overview that consists of manual tasks (square boxes) performed by the policy analyst (white boxes) or crowd workers (shaded boxes) and automated steps performed by tools (circles). The arrows point in the direction of data flows, e.g., illustrating where crowd worker annotations are sent to automated tasks. We now discuss each step in more detail.

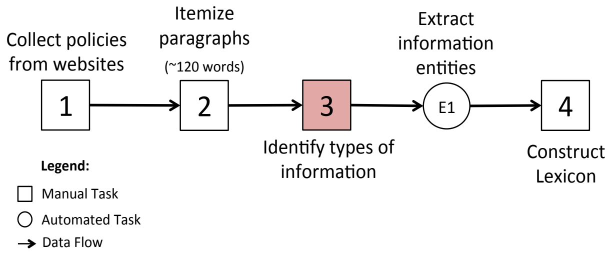
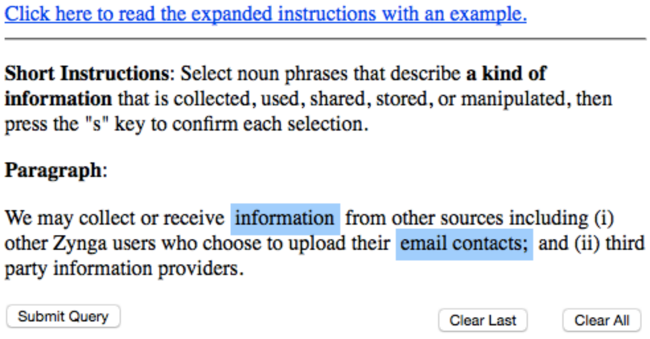


Fig. 1. Entity extraction and lexicon construction workflow

During steps 1 and 2, the analyst prepares the input text to the natural language processing (NLP) tools and crowd worker platform, in this case Amazon Mechanical Turk (AMT). These steps are performed manually by the analyst, once for each policy. For step 1, the input text begins as a text file, which may be extracted from a HTML or PDF file. For step 2, the analyst itemizes the text into paragraphs that average 90-120 words, while ensuring that each paragraph's context remains undivided. This includes ensuring that anaphoric references, such as "it" or "this", are contained in the same paragraph as the noun phrases to which they refer. This invariant can still lead to paragraphs that exceed 120 words, which is balanced by smaller paragraphs 50-60 words. The 120-word average limit determines the average time required by one worker to annotate a paragraph, which we set to 60 seconds. This average time provides workers small, but frequent micro breaks between tasks and it allows workers frequent opportunities to stop annotating text whenever they feel fatigue or boredom.

A. Crowd Worker Micro Tasks

Step 3 is called a crowd worker micro task, because it asks workers to perform a small unit of work, in our case, to annotate noun phrases that correspond to a kind of information, as shown in Figure 2. Following these simple instructions, workers see the ~120-word paragraph and they then may select and annotate relevant phrases using their mouse and keyboard (e.g., the information types "information" and "email contacts"). The results of the task are captured and recorded as part of an AMT batch wherein we ask 5 workers to annotate each paragraph for step 3. This number of workers was determined by prior evaluation [5].


Click here to read the expanded instructions with an example.

Short Instructions: Select noun phrases that describe **a kind of information** that is collected, used, shared, stored, or manipulated, then press the "s" key to confirm each selection.

Paragraph:

We may collect or receive information from other sources including (i) other Zynga users who choose to upload their email contacts; and (ii) third party information providers.

Fig. 2. Crowd worker micro task to annotate information types

The results of steps 3 are then used with the entity extractor in the step E1 to construct the lexicon, which we now discuss.

B. Reusable Lexicon and Entity Extraction

The lexicon is constructed from information type entities, which are unique textual descriptions needed to identify recurring instances of the same concept. For example, the entities in the lexicon should enable us to resolve to a single entity any synonyms, and plural and singular forms of same information type. In step E1 in Figure 1, we perform entity extraction on the annotated noun phrases provided by the crowd workers. These phrases may consist of ambiguous lists and clauses that obfuscate the unique entities and thus pose technical challenges to entity extraction. For example, consider the following privacy policy statement:

"Personal information is *information* that identifies an individual or that can be reasonably associated with a specific person or entity, such as a *name, contact and billing information, Internet (IP) address and information about an individual's purchases and online shopping."*

The statement above has three singleton information type entities (*personal information, information* and *name*); two information type entities with modifiers and a common root word (*contact information* and *billing information*) and one information type entity with parenthesis (Internet Address – IP address). The statement also has two information type entity clauses: *Information about an individual's purchases* and *Information about an individual's online shopping*, which can also be represented as *individual's purchase information* and *individual's online shopping information,* respectively.

Our approach to entity extraction is based on a grounded analysis of 3,850 crowd worker information type annotations from 15 policies. In this analysis, we first examined the sequences of part-of-speech (POS) tags for each annotated phrase to manually identify reliable patterns that we could use to consistently decompose the entity extraction problem into sub-problems (e.g., finding nouns in lists, case-splitting, etc.) Based on our grounded analysis, we arrived at POS-based entity patterns constructed from the POS tags in Table I.

Table II shows a list of information types reported by crowd workers and their corresponding POS-tag patterns based the Stanford POS tagger[1] output; the types are numbered in the first column from 1-6. The types #1 and #6 each show an adjective (JJ), followed by a noun, singular or plural (NN or NNS). Type #4 shows a three-word phrase (an adjective followed by two nouns), whereas type #2 shows a two-word phrase (two proper nouns). We generalized these examples to a well-known regular expression [10] that matches a noun phrase (NP) followed by a clause (CL) as follows:

```
NP=((JJ|RB|VBG|VBD|NN\S?|NN\S?\sPOS)\s)*(NN\S?)
CL= (\s(IN|PRP|TO|VBG|VBN|WDT|WP)\s.*)?
```

TABLE I. Part-of-speech Tags Used by Entity Extractor

Tag	Description
CC	Coordinating conjunction
JJ	Adjective
IN	Preposition or subordinating conjunction
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
POS	Possessive ending
PRP	Personal pronoun
RB	Adverb
TO	*to*
VBG	Verb, gerund or present participle
VBN	Verb, past participle
WDT	Wh-determiner
WP	Wh-pronoun

TABLE II. Annotated information types and POS-tag patterns

#	Annotated Information Type	POS-tag Pattern
1	geographic information	JJ NN
2	local storage	NNP NNP
3	mac address	VBG NN
4	personal health information	JJ NN NN
5	picture	NN
6	pixel tags	JJ NNS

Figure 3 presents the fully automated workflow to extract information type entities that results from the grounded analysis. Because the workflow was developed using grounded analysis, we report the results with each step in the workflow. The workflow begins at "Start", at which point we consider only annotations that two or more workers agreed to. We then test whether the annotation is a list (i.e., does it contains a common list delimiter, such as a comma, semi-colon or POS-tagged coordinating conjunction, or CC tag). If the annotation does not contain a list delimiter, then we test whether the annotation describes a single entity by checking the annotation's POS tag sequence against the NP + CL pattern.

If the NP + CL pattern matches, then we extract a single entity (shaded blue in Figure 3) that we call a ground term. Ground terms are viewed as highly reliable extractions, because they unambiguously match NP + CL pattern and we use these extractions as "ground truth" to disambiguate other steps within the workflow, as we now discuss. Across the 26 policies, we observed that 69.5% of annotations were not lists, and the remaining 30.5% of annotations were lists.

[1] http://nlp.stanford.edu/software/tagger.shtml

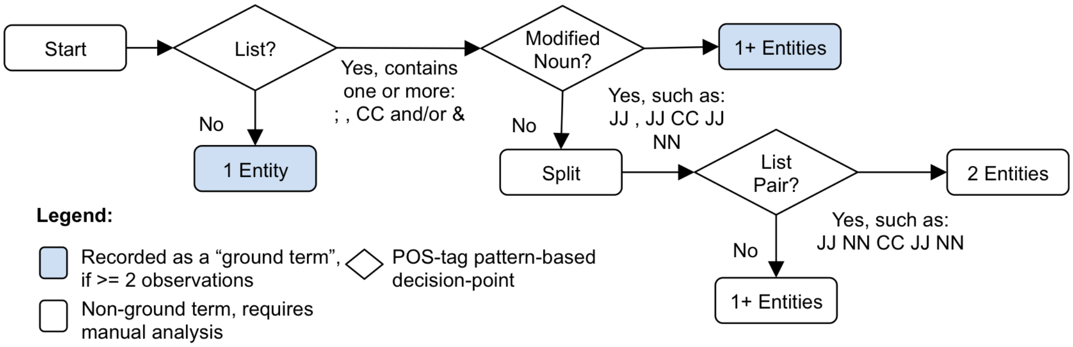

Start
List?
Yes, contains one or more: ; , CC and/or &
No
1 Entity
Modified Noun?
1+ Entities
Yes, such as: JJ , JJ CC JJ NN
No
Split
List Pair?
2 Entities
Yes, such as: JJ NN CC JJ NN
No
1+ Entities
Legend:
Recorded as a “ground term”, if >= 2 observations
Non-ground term, requires manual analysis
POS-tag pattern-based decision-point

Fig. 3. Workflow for the information entity extractor

If the annotation is a list, we first check whether the list is comprised of a list of adjectives preceding a noun phrase. If true, we call this annotation a modified noun, and we extract one or more entities. Modified nouns occurred in 2.5% of all annotations. For these nouns, we performed case-splitting, in which we extracted all permutations of the adjectives preceding the noun. This step includes lists of conjoint adjectives followed by a noun (e.g., “aggregate, statistical information”), as well as disjoint lists (e.g., “geographic and demographic information”). Disjoint lists are split to distribute the modifiers separately across the nouns (e.g., to yield “geographic information” and “demographic information”). Similar to non-list annotations, we view modified nouns as ground terms, or highly reliable extractions. The remaining extractions in the workflow are ambiguous.

In 28% of the annotations, our workflow “splits” the list by delimiter after determining the list is not a modified noun. For each phrase between delimiters, we first check whether the phrase is a list pair by asking, is the phrase comprised of two noun phrases joined by a conjunction? If false, then we check whether each phrase is a previously seen ground term, in which case we extract an entity for that phrase. Alternatively, we store the phrase for manual review by an analyst who determines what entities correspond to the phrase. If the split phrase is a list pair, then we attempt to split the phrase into two entities and check whether both entities are ground terms. If both are not ground terms, then we store the un-split phrase for manual review by the analyst.

Phrases that cannot be automatically resolved by the entity extractor, include vague clauses such as - *information regarding your interaction with the Barnes & Noble enterprise*, which could possibly mean *browsing history*, *clicks*, *cookies*, *website usage information* or another similar information type. This could be attributed to the ambiguity in the privacy policy language, which is difficult to disambiguate even by a human. Similarly the information type annotation “*how you use our mobile applications*” is synonymous to “*mobile usage information*” or one possible interpretation of only “*mobile applications*”.

IV. Results of Scaling Reusable Lexicon

We now present the results from analyzing the 15 privacy policies, which are listed in Table III and covered three domains: news, shopping, and social networking.

TABLE III. List of 15 Privacy Policies Analyzed

Company Name	**Domain**	**Date Acquired**
ABC News	News	09/26/14
Accuweather	News	09/26/14
Amazon	Shopping	02/26/14
Barnes and Noble	Shopping	11/14/14
Bloomberg	News	12/12/14
Costco	Shopping	09/26/14
Facebook	Social Networking	11/14/14
JCPenny	Shopping	12/12/14
Kik	Social Networking	04/17/15
LinkedIn	Social Networking	11/14/14
Reuters	News	12/12/14
SnapChat	Social Networking	04/17/15
Walmart	Shopping	02/25/14
Washington Post	News	12/12/14
WhatsApp	Social Networking	04/17/15

The 15 policies consisted of a total of 3,850 crowd worker annotations. We chose to analyze annotations where two or more crowd workers agreed the highlighted phrase was an information type, which yields 2,270 annotations. Based on this result, a total of 92% of annotations yielded entities. Annotations that did not yield entities would fail to satisfy the workflow and include annotations of verb phrases. Among the 2,198 annotations producing entities, we identified a total of 750 unique entities. Among these, 625 or 83% of the entities were ground terms, which are highly reliable and unambiguous. We identified 276 phrases that required manual intervention, which required less than one hour for an analyst to review and determine the proper entity name for these phrases. Finally, for each policy we observed a minimum of 29% of entities being novel, previously unseen, and a maximum proportion of 62% of entities being novel. We consider this last result more generally in our analysis of the lexicon reusability.

Tables V and VI present an incomplete list of entities that fall under four sensitive information types: contact, financial, personal and technical information. These entity names appeared directly in one or more policies.

TABLE IV. Contact and Financial Information Entities

Contact Information	Financial Information
Address book	Bank routing information
Billing address	Billing payment
Contact list	Credit card security codes
Device's phonebook	Credit history information
Email header information	Credit report
Friend's user IDs	Debit card PINs
Mobile phone number	Financial aid number
Postal address	Order status
Primary email address	Payment information
Screen name	Payment settings
Subscriber information	Purchase history

TABLE V. Personal and Technical Information Entities

Personal Information	Technical Information
Birth date	Browser plug-in versions
Browsing behavior	Clickstream data
Gender	Cookies
Graduation year	Device identifier
Health status	IP address
Job title	Local storage
Language preference	Operating system version
Pharmacist records	Network information
Picture	Search term
Purchasing habits	Signal strength
Precise location	Web addresses

We examined the extent to which the lexicon can predict information types in new privacy policies. This analysis shows that privacy policies have unique entities that are not shared across policies. We define *saturation* to mean the percent reuse of information types in a policy *N* based on the last *N-1* policies previously seen. We counted 100 pseudorandom permutations of the orders of the 15 annotated policies. We observe that near 10 policies, the maximum threshold for saturation of 78% is achieved, meaning, every new policy contributes a sufficient number of unique terms to the lexicon that 22% of the policy terms would not appear in any previously seen policy in the best case, and 73% of the policy terms would be new in the worst case. This best-worst case difference in saturation is determined by which policies had been seen in the *N-1* policies used to compute the ratio. This observation means that the lexicon cannot entirely replace crowd workers, because there appear to always be new terms that the lexicon has never encountered. We plan to scale this study to a larger number of policies to see whether we can reach higher levels of saturation.

V. Discussion and Future Work

We observed that the lexicon reaches a saturation limit of between 31-78% in three domains, which means the lexicon would likely never be deemed complete. That said, we believe the lexicon can still improve NLP analysis of privacy policies by identifying common words and phrases for information types. One question we did not investigate is the extent to which sector-specific subsets of the lexicon are more likely to saturate than the total lexicon with cross-sector terminology.

Our results show that part-of-speech tagging can be used to find over 90% of information entities in annotated texts. Alternatively, other techniques, such as phrase structure grammars that aim to find noun phrases and dependency parsing may be used to yield better performance. In addition, use of machine learning techniques to automate the detection of information types and to extract entities from the identified information types may hold promise. For machine learning, an analyst will need to identify a number of predictive natural language features that suggest where information types are likely to appear. In addition, they will need to develop a larger corpus of correctly annotated documents upon which to train.

Acknowledgment

We thank Hanan Hibshi, Sepideh Ghanavati, Daniel Smullen, Sudhi Wadkar and the CMU RE Lab for their helpful feedback. This research was funded by NSF Award #1330596.

References

[1] Abbot RJ "Program design by informal English descriptions." *Commun. ACM* 26(x11):882–894, 1983.

[2] A.I. Antón, J.B. Earp, "A requirements taxonomy for reducing web site privacy vulnerabilities," *Req'ts Engr. J.*, 9(3):169-185, 2004.

[3] P. André, A. Kittur, S.P. Dow. "Crowd synthesis: extracting categories and clusters from complex data." *17th ACM Conf. Comp. Sup. Coop. Work & Soc. Comp.* 2014. ACM.

[4] D. Berry. Natural language and requirements engineering - nu?, 2001. http://www.ifi.unizh.ch/groups/req/IWRE/papers&presentations/Berry.pdf, accessed 06.09.2015.

[5] T.D. Breaux, F.Schaub, "Scaling requirements extraction to the crowd: experiments on privacy policies," *22nd IEEE Int'l Req'ts Engr. Conf.*, pp. 163-172, 2014.

[6] T.D. Breaux, H. Hibshi, A. Rao. "Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements." *Req'ts Engr. J.*, 19(3): 281-307, 2014.

[7] T.D. Breaux, D. Smullen, H. Hibshi, "Detecting Repurposing and Over-collection in Multi-Party Privacy Requirements Specifications," To Appear: *IEEE 23rd Int'l Conf. Req'ts Engr.*, Ontario, Canada, 2015.

[8] C.A. Brodie, C. Karat, J. Karat, "An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench, Symposium On Usable Privacy and Security (SOUPS), Copyright IBM Corp, 2006.

[9] A.. Dardenne, S. Fickas, A. van Lamsweerde. "Goal–directed requirements acquisition," *Sci. Comp. Prog.*, 20:3-50, 1993.

[10] J.S. Justeson, S.M. Katz, 1995. “Technical terminology: some linguistic properties and an algorithm for identification in text.” *Natural Lang. Engr.* 1:9-27.

[11] S. L. Lim, A. Finkelstein. “StakeRare: using social networks and collaborative filtering for large-scale requirements elicitation,” *IEEE Trans. Soft. Engr.,* 38(3): 707-735, 2012.

[12] A.K. Massey, J. Eisenstein, A.I. Anton, P.P. Swire, "Automated text mining for requirements analysis of policy documents," *21rd Int’l Conf. Req’ts Engr.*, pp. 4-13, 2013.

[13] A. McDonald, L.F. Cranor. “The Cost of Reading Privacy Policies”, *I/S: A Journal of Law and Policy for the Information Society*, 2008.

[14] G.R. Milne, M.J. Culnan, H. Greene “A Longitudinal Assessment of Online Privacy Notice Readability.” *J. Public Policy & Marketing*: 2006, 25(2): 238-249.

[15] A. J. Quinn, B. B. Bederson. “Human Computation: A Survey and Taxonomy of a Growing Field.” *Conf. Hum. Factors in Comp. Sys.,* pp. 1403–12, 2011, ACM.

[16] R. Snow, B. O’Connor, D. Jurafsky, A. Y. Ng. “Cheap and Fast—but Is It Good?: Evaluating Non-Expert Annotations for Natural Language Tasks.” *Conf. Emp. Meths. in NLP*, pp. 254–63, 2008, ACL.

[17] S. Zimmeck, S.M. Bellovin. “Privee: An Architecture for Automatically Analyzing Web Privacy Policies.” *USENIX Security*, pp. 1-16, 2014.