Beyond The Text: Analysis of Privacy Statements through Syntactic and Semantic Role Labeling

Yan Shvartzshnaider
York University, Toronto, Canada
yansh@yorku.ca

Ananth Balashankar[*]
New York University, New York, USA
ananth@nyu.edu

Thomas Wies
New York University, New York, USA
wies@cs.nyu.edu

Lakshminarayanan Subramanian
New York University, New York, USA
lakshmi@nyu.edu

Abstract

This paper formulates a new task of extracting privacy parameters from a privacy policy, through the lens of Contextual Integrity (CI), an established social theory framework for reasoning about privacy norms. Through extensive experiments, we further show that incorporating CI-based domain-specific knowledge into a BERT-based SRL model results in the highest precision and recall, achieving an F1 score of 84%. With our work, we would like to motivate new research in building NLP applications for the privacy domain.

1 Introduction

A privacy policy informs users about a company's information handling practices. However, privacy policies are lengthy documents, full of incomplete and vague statements that impose a significant cognitive burden on the reader to infer whether a given service respects their privacy (Bhatia et al., 2016a; Bhatia and Breaux, 2018; Reidenberg et al., 2015).

This challenge has inspired many recent works in applying natural language processing and machine learning techniques to automatically process privacy policies and retrieve the relevant information (Harkous et al., 2018; Ravichander et al., 2019). While these efforts help in identifying paragraphs in the privacy policy that mention sensitive information (Evans et al., 2017; Bhatia and Breaux, 2015), opt-out clauses (Sathyendra et al., 2016) or description of data collection practice (Sadeh et al., 2014), they focus on the policy as a whole rather than on the privacy implication of the individual privacy statements that it contains. In particular, they do not aim to identify relevant and often missing contextual information that is critical for unambiguously understanding the scope of individual statements. This paper focuses on a new NLP task that aids the analysis of privacy policies at this more fine-grained level.

To illustrate the problem, consider a typical example of an ambiguous privacy statement: "**Yahoo** *collects* **information about your transactions with us and with some of our business partners**, *including* **information about your use of financial products and services that we offer.**" At first glance, the statement may seem to provide all the relevant information about a first-party collection of transactional data. However, it in fact misses some crucial contextual information. To understand what is missing, we use the contextual integrity (CI) framework (Nissenbaum, 2009). CI defines privacy as an appropriate flow of information which is expressed in terms of five essential *CI parameters*: Sender, Recipient, Subject, Information Type, and Transmission Principle. The latter is a constraint on the information flow expressing the condition under which information is being transferred. The statement in our example specifies only 3 out of the 5 necessary parameters (highlighted in bold) – Subject, Recipient, and Information Type. This leaves the sender of the information and transmission principle to the reader's interpretation. In some cases, the relevant missing information appears in different places in the policy, for example, under different sections such as "When do we collect your information" or "Our partners". These, however, do not help in contextually positioning the above statement so that the reader can determine whether their expectations have been met.

In this paper, we show that existing NLP models and techniques can assist a human annotator in identifying relevant privacy parameters in the policy text. The proposed novel NLP task can support the following applications: automate comparative analysis of privacy policies using the theory of contextual integrity (Shvartzshnaider et al., 2019a; Sanfilippo et al., 2019); extraction and enforcement of prescribed CI-based policy from legal and co-

[*] currently at Google

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 85–98
December 7, 2023 ©2023 Association for Computational Linguistics

operate document (Shvartzshnaider et al., 2019a; Sanfilippo et al., 2019), an enhanced auditing of existing privacy policies for correctness and consistency (Andow et al., 2019).

2 Contextual Integrity Primer

The theory of Contextual Integrity (CI) defines privacy as appropriate information flow in accordance with governing privacy norms. As mentioned in the introduction, CI provides a framework to capture and compare information flows against established norms. To perform an analysis of the privacy implications of a given information flow, CI requires identifying five essential parameters: actors (sender, receiver, subject), the type of information (attribute), and the condition or purpose of the information exchange (transmission principle). It is critical to state all five parameters to ensure a non-ambiguous privacy implication analysis. A misalignment of parameter values–i.e., an information flow that deviates from the established norm–constitutes a potential privacy violation.

For example, consider a medical context, an established norm states that a patient (sender) could share their (subject) medical information (information type) with their doctor (receiver) confidentially (transmission principle). An information flow (e.g., generated by an app or a system) that deviates from this norm (e.g., by changing the value of TP to public and/or sending the medical information to a different receiver) potentially violates the patient's privacy and requires further investigation. CI provides a heuristic to examine violating information flows in terms of how they contribute to values, functions, and purposes of a particular context (Nissenbaum, 2009).

CI and Privacy Policies By identifying the values of CI parameters in a privacy statement, we can reason about privacy implications in a way that more closely aligns with users' privacy expectations and societal norms.

This analysis can help in identifying potentially confusing or misleading statements, e.g., when one of the five parameters such as transmission principle or receiver is missing or ambiguous (Shvartzshnaider et al., 2019a). Furthermore, one can use the identified parameters to formalize the expressed informational norms and privacy rules in formal logic (Shvartzshnaider et al., 2019b; Datta et al., 2011). These formalisms can in turn be used to build systems that enforce the specified rules or automatically audit information flows to detect rule violations.

3 Related Work

Several recent efforts have focused on identifying important and relevant privacy statements using constituency parsing (Sathyendra et al., 2017, 2016; Evans et al., 2017), logistic regression (Ammar et al., 2012) and crowdsourcing (Wilson et al., 2016b) techniques. These works focus on other aspects of privacy policy analysis such as opt-out disclosure, right to information access, etc.

As we discuss in Section 4, our work explicitly looks to map the privacy statement to a fixed set of parameters. We also show that Question Answering (QA) models do not perform satisfactorily when applied to our task. Similar limitations of the reading comprehension models were observed by Ravichander et al. (2019), who composed the PRIVACYQA dataset, an annotated corpus consisting of 1750 questions about the contents of privacy policies such as "What data does this game collect?" and "Will my data be sold to advertisers?".

In prior work on automatic privacy statement analysis, Bhatia et al. (2016b) extracted privacy statements on information handling practices such as "collecting your e-mail address" or "sharing your location" using a typed dependency parser and crowdworker annotations. More relevant to our efforts, Bhatia and Breaux (2018) applied Semantic Roles theory to manually annotate five privacy statements and identify action verbs (action data) such as "collection", "retain", "use", "transfer" and associated semantic roles that capture who performs the action, how the action is carried out, etc.

Andow et al. (2019) developed PolicyLint to analyze privacy policy using sentence-level NLP techniques to capture statements in a four-element tuple (actor, action, data object, entity) to identify contradictory statements. Andow et al. (2020) used a simplified 3-tuple statement abstraction (actor, data object, entity), and identify incorrect and ambiguous statements. Trimananda et al. (2022) build on this effort to audit the traffic and privacy policies in Oculus VR systems. Per CI theory, both OVRseen and PoliCheck capture insufficient information to perform a privacy implication analysis, which requires five CI parameters. Nevertheless, we can use PolicyLint and PoliCheck libraries to help process privacy policies to generate statements that we can feed into our annotation pipeline.

Shvartzshnaider et al. (2019a) crowdsourced privacy policies annotation to compare policy versions, identifying missing contextual information and overloading of parameters that contribute to users' inability to understand the prescribed information practices. Our work automates the task of annotating privacy policies with the CI parameters.

4 Task Formulation

The CI parameter extraction task is as follows. Given a privacy statement $stmt$, apply a mapping function M to extract the CI parameters: sender, receiver (r), subject (s), attribute (att), transmission principle (tp):

$$M(stmt) = (s, r, sub, att, tp)$$

The main challenge behind the task is identifying the lexical items in the statement that correspond to the contextually relevant values to help downstream NLP tasks perform the privacy analysis. This is not a trivial task as privacy policies are not written with CI in mind. Often, policies are written by legal and policy teams whose primary concern is not readability. Many privacy statements are missing essential CI parameters and often comprise syntactically complex sentences (Bhatia and Breaux, 2018).

In the absence of an automatic way to extract CI parameters, researchers have employed crowdsourcing and manual annotation to perform the analysis (Shvartzshnaider et al., 2019a). However, these methods do not scale due to the high cost of annotation by experts and hence we propose an ML-based approach by training or mapping existing models trained on NLP tasks to extract CI parameters. Further, we expand on a growing body of research in human-in-the-loop ML-assisted validations (Section 5.3) to evaluate the precision and recall of the model annotations.

5 Methods

In this section, we describe our method to perform the CI parameter extraction task by incorporating the contextual semantics of CI parameters to modify two conventional NLP techniques: Syntactic Dependency Parser (DP) and BERT-based Semantic Role Labeling (SRL). We then describe how these individual techniques can be integrated into an end-to-end model to extract CI parameters from privacy policies. In addition, in our evaluation (Section 6), we also compare our method to three baseline models tailored to address the CI parameter extraction task: Question-Answering, BERT, and Hidden Markov Model.

5.1 Enhancing NLP tasks using CI

We focus on Syntactic DP and SRL-based approaches and describe how we have incorporated CI-based domain-specific knowledge.

5.1.1 CI-based Dependency parsing

Dependency parsing is the task of identifying syntactic roles or dependency types for each of the words in a sentence. This involves parsing a sentence and identifying the syntactic structure denoting the grammatical rules that govern a language. The parser (Honnibal and Johnson, 2015) uses a non-monotonic transition system that allows for a large number of parse trees for each intermediate state. This modified arc-eager transition system has been shown to repair earlier parsing mistakes. It allows for overwriting previous parsing decisions and achieves a higher accuracy as compared to the greedy parsing approaches. Not all the dependency types identified for the English language are relevant in our study. We use the DP outputs to identify the relevant CI parameters in the privacy statement.

To identify CI parameters at a single sentence level using local relationships, we run a typed dependency parser (DP) on the text of the policies. We accept paragraphs as input, split them into sentences, and parse each sentence using the Spacy I/O[1] dependency parser. The model (Honnibal and Montani, 2018) achieves near state-of-the-art performance on most NLP tasks[2]. Based on our analysis of DP outputs for a sample of statements, we identified the dependency types that mapped to specific CI parameters as shown in Table 1.

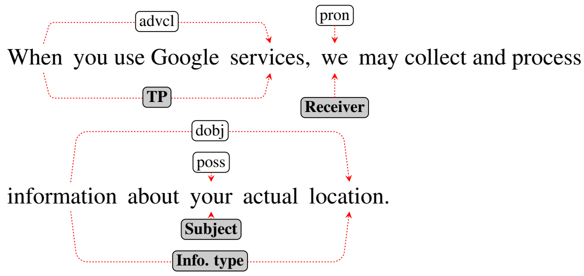
CI Parameter Type	**Dependency types**
Attribute	*dobj*, *parataxis*, *nsubj-pass*
Sender/Receiver	*nsubj*, pronouns
Transmission Principle	*xcomp*, *ccomp*, *advcl*, *oprd*
Subject	*poss*, *agent*

Table 1: Mapping of dependency types corresponding to CI parameters. To represent dependencies we use the Stanford Typed Dependency Manual (De Marneffe and Manning, 2011) notations.

[1] https://spacy.io

[2] https://spacy.io/usage/facts-figures

For example, for the following statement from the Google privacy policy, the DP parser will return the following dependency type tags (white nodes), which are mapped to the corresponding CI parameter (gray nodes):



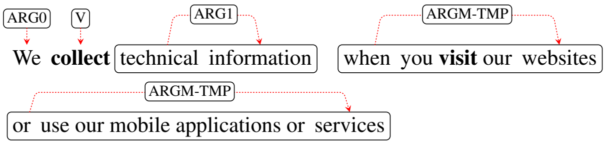
advcl
pron
When you use Google services, we may collect and process
TP
Receiver
dobj
poss
information about your actual location.
Subject
Info. type

Note that, as is evident in Table 1, the dependency types cannot distinguish between the parameter of sender and receiver. For this, we defer to the task of SRL to identify based on the semantic meaning of the word.

CI-based Semantic Role Labeling Semantic Role Labeling (SRL) is the task of mapping words or phrases in a sentence to a semantic role such as that of an agent, goal, or result (Jurafsky and Martin, 2014). Often, in the classic natural language processing pipeline, this task is considered to have subsumed syntactic and parts-of-speech tasks within it (Tenney et al., 2019). For example, the task of distinguishing between a sender and a receiver can be done through SRL, but not through syntactic DP.

Similar to DP, we map the semantic roles to the relevant CI parameters. Table 2 shows the CI parameter mapping based on a verb's syntactic arguments. For example the verb "collect" has the following associated arguments (see PropBank corpus (Martha et al., 2005)): `ARG0`: agent, entity acquiring something, `ARG1`: thing acquired, `ARG2`: source, `ARG3`: more specific attribute of `ARG1` being collected, `ARG4`: benefactive.

To recover the predicate argument structure of a sentence we use an out-of-the-box AllenNLP implementation of the Bidirectional LSTM and SRL BERT models (Stanovsky et al., 2018; Shi and Lin, 2019). The first approach uses a bidirectional LSTM model for sequence tagging that allows for multiple overlapping tuples per sentence by extending deep BIO taggers for semantic role labeling, trained on the Open information extraction task. The second approach uses two sequence-to-sequence BERT models, one for predicate sense disambiguation, and another for argument identification and classification, with relevant parts of the sentence and predicate as part of the input, with a sequence of semantic role labels for outputs. For example, for the following statement, the BERT SRL model returns:



ARG0
V
ARG1
ARGM-TMP
We collect technical information
when you visit our websites
ARGM-TMP
or use our mobile applications or services

Example 1: SRL processed statement from a Walmart privacy policy

Based on the syntactic analysis of privacy policy statement sentences, we mapped the arguments onto the CI parameters. In the above example, `ARG0` is mapped to Recipient. `ARG1` is an Attribute, and `ARGM-TMP` is the TP. For each of the verbs, these mappings are slightly different, as shown in Table 2. This approach, although crude, covers a significant class of privacy policy statements that prescribe the flow of information.

	"Sending" action verbs	**"Receiving" action verbs**
Sender	*ARG2*	*ARG0*
Receiver	*ARG0*	*ARG2*
Attribute	*ARG1*, *C-ARG1*	
TP	*ARGM-TMP*, *ARGM-ADV*, *ARGM-MNR* *ARGM-PNC*, *ARGM-CAU*	

Table 2: Mapping semantic roles (notations) to specific CI parameters.

5.1.2 CI-related Semantic Frames

The SRL model returns verb-argument predicates for all the identified verbs in a sentence. Some of these verbs are not relevant to information exchange. For example, in the statement in Example 1, the verb "visits" does not convey semantically meaningful information regarding the exchange of technical information.

To reduce the number of false positives, we provide a list of verbs to the algorithm which highly correlates with information exchanges. The choice of verbs is based on a frequency analysis of verbs, along with filtering out verbs that are not related to information flows (as shown in Table 11 in the Appendix). It is helpful to think of this approach through the lens of the linguistic theory of Frame semantics (Fillmore et al., 1976), which posits that the specific meaning of words (frame elements) can be understood only as part of a particular context (semantic frames). In our approach, we would invoke CI-related semantic frames. Specifically, we look for SRL-predicates that are associated with any transfer of information (actual or perceived). This includes a list of verbs such as “sending”, “sharing”, “transmitting” and others that frequently appear in privacy policy text as identified by our POS analysis of the OPP-115 policies. In addition to invoking a general semantic frame, we differentiate between different roles of associated argument with each predicate. In particular, for predicates like “sending”, “sharing”, “transmitting” the ARG2 is typically associated with the agent role of a “sender”, the ARG1 captures what was “sent” and ARG0 is associated with the receiving agent role. For verbs like “gather”, “collect”, “receive”, “acquire” the roles are reversed: ARG0 is typically associated with a “sending” agent role, the ARG1 describes what is “Received”, and ARG2 is associated with the “receiving” agent role. Grouping the verbs signifying a “sending” or a “receiving” action helps us map the corresponding arguments to the relevant CI parameters for Senders and Receivers. The mapping for TP and Attribute remains the same for all verbs. Finally, our SRL mapping does not include a semantic role mapping of the Subject parameter. We operate on the assumption that the CI subject parameter in most statements is the user.

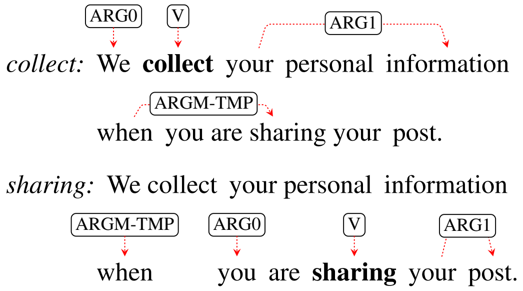
5.1.3 Clues from CI to Improve SRL

We first gather the arguments from the SRL model and map them to CI parameters, as shown in Table 2. Identifying the arguments for all verbs in the privacy statement results in high recall numbers. Nevertheless, the precision suffers because not all of the verbs need to be invoked. To reduce the number of false positive mappings, we implement an Algorithm 1 which analyzes all the relevant SRL verbs to check whether any of them appear as part of the Transmission Principle (TP) relative to another verb. Specifically, after the arguments are mapped to CI parameters, we iterate through each verb in the sentence and remove it if it has already been mapped as part of a TP for another verb.

Algorithm 1

```
$Args \leftarrow SRL(sentences)$
$Dict(Verb, CI_{params}) \leftarrow MapToCI(Args)$
for all $verb_i \in Dict(Verb : CI_{params})$ do
  for all $verb_j \in Dict(Verb : CI_{params})$ do
    if $verb_i \in Dict[verb_j][TP]$ then
      $RemoveMappings(Dict(verb_i))$
    end if
  end for
end for
```

For example, in the following statement, the SRL model will pick up two predicates (verbs) and corresponding arguments:


ARG0
V
ARG1
collect: We collect your personal information
ARGM-TMP
when you are sharing your post.
sharing: We collect your personal information
ARGM-TMP
ARG0
V
ARG1
when you are sharing your post.

These arguments will be mapped to CI parameters, as described in the previous section. The verb “share" is redundant in this context since it is part of the TP of the verb “collect." Once we identify the *redundant* verb, we ignore all arguments associated with it, i.e., our algorithm does not consider these results. We do keep those parameters that overlap with the parameters produced by non-redundant verbs. For instance, in our example, we ignore the verb “share" and the associated with it arguments. Specifically, the [**ARG0:** you] and the [**ARG1:** your post] which otherwise will be mapped to a CI sender and attribute parameters, respectively.

5.2 Expert Annotation Task

We perform *m*odel-assisted annotation of close to a thousand privacy statements that discuss information exchanges across 50 policies (see Table 9 in the Appendix) from the OPP-115 Corpus (Wilson et al., 2016a).

5.2.1 Data processing

To extract relevant statements, we relied on existing OPP-115 corpus annotations of “data practices” in each of the segments of the policy. We limit our CI parameter extraction to labeled segments of the policy that discuss information exchanges such as segments labeled as “First Party Collection/Use”, “Third party sharing/collection”, and “Data Retention". We also cleaned that data to remove short parts of speech that are intrinsically captured by a syntactic parser.

The selected statements from the privacy policies were then presented to a human annotator, one of the authors who is an expert on CI. The expert then marked the valid results for each of the privacy statement sentences and CI parameters, which form the ground truth. A sentence was marked as a "valid" flow if it prescribed an information exchange of any kind. Otherwise, by default, all sentences are considered "invalid." Overall, the extraction phase resulted in a total of **2808** privacy statement sentences, out of which **994** were labeled as valid, containing **4048 CI parameters**. On average, a policy contains 18 valid statements, with outliers of 4 and 45 valid statements. Furthermore, our dataset contains: a) **4333** SRL labels (2846 true positives and 1487 false positives), **5974** DP labels (2484 true positives and 3490 false positives). Table 8 in the Appendix shows the annotation labels breakdown for each CI parameter.

Annotation reliability We assessed the reliability of the expert by co-coding with another author annotating 10 of the 50 policies to achieve a substantial to excellent agreement with Cohen's Kappa score range of 0.67-0.88 on the CI parameter annotations. Table 7 shows the corresponding breakdown of Cohen's Kappa score (McHugh, 2012) for each of the CI parameters.

Parameter	**Cohen's Kappa Agreement score**
Sender	0.88
Receiver	0.80
Attribute	0.63
Subject	0.73
TP	0.68

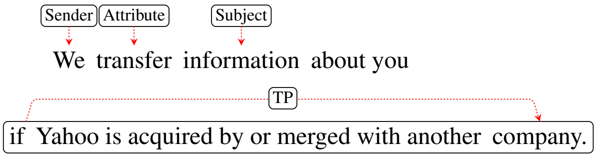
Table 3: The Cohen's Kappa agreement scores for CI parameters annotation.

5.3 ML-assisted expert validations

The outputs of the methods we have proposed in Section 5 may have an overlap with the ground truth. However, for CI parameters like the transmission principle, it is not clear if relying on an exact string match or a fuzzy match would consistently capture what makes any output a transmission principle. Hence, comparing the outputs with the expert annotations requires a careful analysis that goes beyond unigram/bigram/trigram match or fuzzy match of the two spans of text. Instead, we perform a validation of the outputs generated by the model, where the expert can confirm if the output matches the expectation and criteria for the five CI parameters[3]. Such a human-in-the-loop annotation process is necessary to check for any inconsistencies as shown in prior work for image annotations (Zurowietz et al., 2018; Marques and Barman, 2003), knowledge-oriented text tasks (Klie et al., 2018), clinical text annotations (South et al., 2012). The task involved comparing against ground-truth human annotations, and selecting among the model-based outputs, the ones that match the definition of a contextual integrity parameter. For example, although the model generated output "has an xx% match as per fuzzy match with the ground truth parameter:", it still does not capture the critical component that defines the transmission principle. Similarly, we see patterns for other CI parameters that are not easily definable in terms of rules that can scale with the number of models we compare against. This limitation is one that is highlighted in human-in-the-loop annotations and validations with domain expertise (Monarch, 2021), and improving the scaling of this process is left for future work.

5.4 Baseline Models

Question Answering As an exploratory experiment, we used an open domain QA model (AllenNLP library with the BiDAF model with GloVe embeddings[4]) to answer CI-related questions regarding each privacy statement. For example, to identify the marked CI parameters in the following privacy statement, we asked a variation of these questions: *"Who is transferring?"*, *"What is being transferred?"*, *"Who is the subject?"*, *"Who is the receiver/recipient?"*, *"Why, When, and How is the transfer facilitated?"*:



We transfer information about you

if Yahoo is acquired by or merged with another company.

BERT We frame the CI parameter extraction task as a sequence-to-sequence transformation problem

[3] See Figure 3 in the Appendix of a screenshot of the choices that the annotation tool built for this purpose

[4] bidaf-model-2017.09.15-charpad.tar.gz

to fine-tune a Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018) on our dataset to map a sequence of words in privacy statements to a corresponding sequence of CI tags. The input sentence is tokenized as a sequence, and the output sequence is a per-token mapping between {None, Sender, Receiver, Attribute, TP and Subject} parameters. For training and testing, we transformed our dataset into the CoNLL2003 format and used the AllenNLP re-implementation of (Gardner et al., 2018) with the train-test split ratio as 80/20 and values of hyperparameters taken from (Gardner et al., 2017).

Hidden Markov Model As a baseline, we also formulate the CI parameter extraction as a part-of-speech (POS) tagging task and use a Hidden Markov Model (HMM) probabilistic model (Jurafsky and Martin, 2014) for annotating words in a sentence. Specifically, we train a trigram HMM by converting the dataset to CoNLL-2003 format (Sang and De Meulder, 2003) with CI parameters as the target labels and an 80/20 train-test split. By default, HMM relies on the Markov assumption that the probability of a particular state only depends on the preceding state. However, in order to enrich our HMM model, we consider the two previous states when predicting the current CI parameter, turning it into a trigram model. Further, we obtain the final transition probability distribution by linearly combining unigram, bigram, and trigram probability distributions:

$$P(t_i \mid t_{i-1}, t_{i-2}) = \lambda_1 P(t_i \mid t_{i-1}, t_{i-2}) + \\ \lambda_2 P(t_i \mid t_{i-1}) + (1 - \lambda_1 - \lambda_2) P(t_i)$$

The parameters λ_1 and λ_2 are fine-tuned on the validation set with values 0.42 and 0.48 providing the best results. The Viterbi algorithm (Forney, 1973) is used in the decoding phase for the extended model.

6 Evaluation

We now compare the performance of the different models on the CI Annotation task using the dataset in Section 5.2 with some common baseline models repurposed for CI.

6.1 Baseline Performance

Table 4 shows the results of our expeditionary experiment. The overall F1 scores for the QA model indicate poor results for the extraction of all CI parameters. In our experiment, QA outputs multiple phrase predictions for each of the parameters. For precision, we calculate true positives as a fraction of all positives predicted for each parameter. For recall, we calculate the fraction of true positives to all correct parameters.

	Recall	**Precision**	**F1**
Attribute	0.21	0.15	0.17
Receiver	0.08	0.06	0.07
Sender	0.03	0.02	0.02
Subject	0.07	0.02	0.03
TP	0.21	0.16	0.18

Table 4: Precision, Recall, and overall F1 score for QA Comprehension model used for the CI parameter extraction task. The recall and precision values for a parameter are calculated by macro averaging over privacy statements.

This result aligns with previous uses of QA in the privacy domain (Ravichander et al., 2019), which observed that compared to a human annotator, using standard reading comprehension models for privacy policies returns unsatisfactory results. These experiences suggest that QA models require additional heuristics to filter the many false positives as a result of operating on a paragraph level and not on sentence-level statements.

Table 5 shows the results of training a trigram Hidden Markov Model and a fully-supervised BERT for the CI-parameter extraction task. Both models perform relatively poorly for our task, especially when it comes to the "Sender" parameter. HMM's overall F1 scores are slightly better for detecting other parameters, with the highest F1 score achieved for the TP parameter in both models.

	Recall		**Precision**		**F1**	
CI Param.	HMM	BERT	HMM	BERT	HMM	BERT
Attribute	0.67	0.61	0.57	0.51	0.62	0.55
Receiver	0.44	0.53	0.52	0.36	0.47	0.43
Sender	0.08	0.12	0.15	0.16	0.10	0.14
TP	0.80	0.75	0.68	0.55	0.74	0.64

Table 5: F1 Scores for fully-supervised HMM and fine-tuned BERT model. The recall and precision values are calculated on word level over the whole test set.

6.2 CI Improved Performance

Table 6 shows precision and recall for both DP (Spacy (Honnibal and Montani, 2018)) and SRL

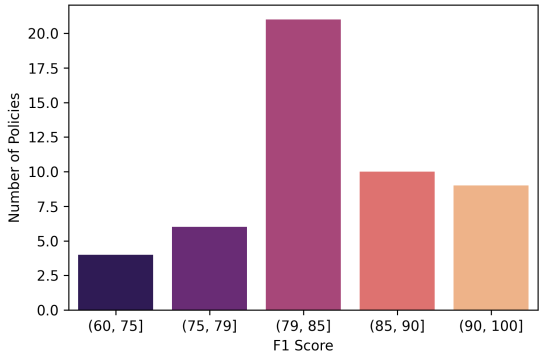


Figure 2: Histogram of F1 scores across privacy policies

models (AllenNLP BERT SRL (0.8.5) and (2.1.0)). Both DP and the two SRL models have high recall numbers. However, in DP the precision is lower, indicating that although DP identifies all the relevant instances, it also produces a large number of false positives. Overall, compared to DP, both SRL models have higher precision and recall. The AllenNLP (2.1.0) SRL-adapted structured BERT model (Shi and Lin, 2019), with comparable results[5] to robust models like RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2019) on CoNLL SRL datasets (Pradhan et al., 2013), produced better results compared to the earlier version (0.8.5) (Stanovsky et al., 2018). Furthermore, compared to DP, the SRL models have slightly higher recall numbers and much higher precision. We, however, note that SRL did not process 49 valid statements missing 159 valid CI parameters. In this case, the statements were ignored by our SRL algorithm because the semantic frame was not triggered as they contained verbs that our algorithm did not track. Some of these verbs are not always associated with information exchange, like "sell" and "rent".

6.2.1 Improved SRL

Table 6 shows the results for SRL after applying our algorithm incorporating domain-specific heuristics on both BERT SRL model versions. The precision results have improved across all the parameters, affecting recall only slightly. We note that our F1 metric is calculated on the phrase prediction level.

Performance across Policies: Figure 2 shows F1 score distributions for the annotated policies. The majority of policies (31) have F1 scores in the range of 80-89, which is consistent with the average F1 scores per parameter. Nine policies perform much better giving a high F1 value of more than 90, six policies are within the 75-80 range, and four policies in the 60-75 F1 range. Refer to Table 9 in the Appendix for the F1 scores for each policy.

Model	**CI Parameters**	**Recall**	**Precision**	**F1**
DP	Attribute	0.68	0.43	0.53
	Subject	0.79	0.26	0.39
	TP	0.76	0.61	0.68
SRL (0.8.5)	Attribute	0.92	0.70	0.80
	Receiver	0.94	0.74	0.83
	Sender	0.95	0.63	0.75
	TP	0.89	0.69	0.78
SRL (2.1)	Attribute	0.88	0.71	0.79
	Receiver	0.88	0.75	0.81
	Sender	0.90	0.64	0.75
	TP	0.93	0.70	0.80
CI-SRL (0.8.5)	Attribute	0.90	0.75	0.82
	Receiver	0.88	0.79	0.83
	Sender	0.90	0.73	0.81
	TP	0.88	0.81	0.84
CI-SRL (2.1)	Attribute	0.88	0.78	0.83
	Receiver	0.87	0.81	0.84
	Sender	0.88	0.76	0.82
	TP	0.84	0.83	0.84

Table 6: F1 Scores for all the models: DP, SRL models and Improved SRL (CI-SRL). The recall and precision values for each parameter are calculated by macro averaging over privacy statements.

We analyzed the privacy statements for which our heuristic algorithm achieved low precision scores to understand the reasons behind the poor performance. Our SRL-based algorithm performed poorly on statements with semantically complex sentences. Semantically complex statements comprise multiple verb-predicates with related arguments that result in a large number of false positives. We also noticed a large number of false positives in long connected sentences that due to improper punctuation appear as a single sentence to our algorithm.

These cases are not only problematic for an NLP task but also require significant cognitive effort to analyze the privacy implications of the pre-

[5] https://paperswithcode.com/sota/coreference-resolution-on-conll-2012

scribed information flows. Rather than adapting our method to yield better results in these cases, we can use it to detect these complex sentences so that they can be restated more clearly.

7 Discussion

In this paper, we describe an NLP-based method to perform automatic CI annotation of privacy statements. This work has several implications for different stakeholders, some of which were discussed in (Shvartzshnaider et al., 2019a).

For consumers, who often find privacy policies too lengthy and complex to read, the CI annotation provides the ability to query the privacy policies for specific flows it prescribes. These can involve asking about mentions of specific "senders" and "recipients" in the privacy policies, the type of information being collected or shared, and listing the conditions and purposes. Beyond enumerating the prescribed CI parameters, the method enables contextual queries that reflect up to five CI parameters. For example, instead of asking about mentions of "location information," the query can include the senders, and recipients of interest, such as third parties, and contractors. The query can be further refined to include specific constraints under which location data is transferred, e.g., when the user is using the app.

The CI abstraction also allows for querying different privacy policies, potentially, from different sectors. Moreover, it can serve as a way to compare different versions of the same privacy policy to examine the new changes that were introduced in the latest version. We can identify which new flows were added or removed and, consequently, what data types, senders, recipients, subjects, and conditions were changed.

Finally, combining with existing crowdsourcing methodologies of learning users' privacy expectations (Apthorpe et al., 2018; Shvartzshnaider et al., 2016; Apthorpe et al., 2019), we can perform a large-scale analysis of what privacy policies align with users' expectations and existing societal norms.

For relevant agencies like the Federal Trade Commission, the CI-based analysis of privacy policies can serve as a robust auditing technique for prescribed data handling practices. Using the CI framework, the resulting information flows can be automatically compared with existing regulation (Apthorpe et al., 2019) and users' privacy expectations (Shvartzshnaider et al., 2016; Apthorpe et al., 2018). Furthermore, the analysis can help detect privacy statements with missing contextual information that results in ambiguous flows. Ultimately, we envision our work used as part of the overall evaluation framework for performing dynamic and static analysis of digital services, similar to (Sanfilippo et al., 2019). Specifically, the CI annotated policy fed into an automatic test suite for possible violation of privacy policy, regulation, or societal expectations.

7.1 Limitations

As discussed in Section 6, our methods result in high F1 scores (>80%) of annotations. Nevertheless, there are cases where our approach does not perform as well. Our evaluation of different models shows that a better SRL model coupled with domain knowledge heuristic improves results. Previous results (Shvartzshnaider et al., 2019a) showed that crowdworkers are able to identify CI parameters with high precision. Coupled with our approach, crowdworkers can filter the false positives and increase the precision further. Eliminating the need for validation of model outputs in the evaluation phase would also make it easy to evaluate better models in the future.

8 Conclusion

In this paper, we formulate the new CI parameter extraction NLP task for the analysis of privacy statements. We adapt several conventional NLP models (QA, HMM, BERT, DP and SRL) to perform the task and demonstrate that it cannot be solved trivially. In our evaluation of privacy statements of 50 real-world privacy policies, we show that a method combining clues from CI with syntactic DP coupled with type-specific SRL obtains the highest F1 score. We build on this insight to devise an algorithm that incorporates CI-based domain-specific knowledge to achieve higher precision and recall. The proposed algorithm post-processes ML outputs and increases automation of a tedious task that has so far been performed manually. Further improvements of this task, leveraging domain knowledge for complex scenarios will directly benefit downstream applications ranging from aiding the design and analysis of privacy policies to building systems that meet users' privacy expectations by construction.

References

Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic categorization of privacy policies: A pilot study.

Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. Policylint: investigating internal privacy policy contradictions on google play. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 585–602.

Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words: Entity-Sensitive privacy policy and data flow analysis with PoliCheck. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 985–1002. USENIX Association.

Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–23.

Noah Apthorpe, Sarah Varghese, and Nick Feamster. 2019. Evaluating the contextual integrity of privacy regulation: Parents' iot toy privacy norms versus {COPPA}. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 123–140.

Jaspreet Bhatia and Travis D Breaux. 2015. Towards an information type lexicon for privacy policies. In *Requirements Engineering and Law (RELAW), 2015 IEEE Eighth International Workshop on*, pages 19–24. IEEE.

Jaspreet Bhatia and Travis D Breaux. 2018. Semantic incompleteness in privacy policy goals. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 159–169. IEEE.

Jaspreet Bhatia, Travis D Breaux, Joel R Reidenberg, and Thomas B Norton. 2016a. A theory of vagueness and privacy risk perception. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*, pages 26–35. IEEE.

Jaspreet Bhatia, Travis D Breaux, and Florian Schaub. 2016b. Mining privacy goals from privacy policies using hybridized task recomposition. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 25(3):22.

Anupam Datta, Jeremiah Blocki, Nicolas Christin, Henry DeYoung, Deepak Garg, Limin Jia, Dilsun Kaynar, and Arunesh Sinha. 2011. Understanding and protecting privacy: Formal semantics and principled audit mechanisms. In *International Conference on Information Systems Security*, pages 1–27. Springer.

Marie-Catherine De Marneffe and Christopher D Manning. 2011. Stanford typed dependencies manual.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Morgan C Evans, Jaspreet Bhatia, Sudarshan Wadkar, and Travis D Breaux. 2017. An evaluation of constituency-based hyponymy extraction from privacy policies. In *Requirements Engineering Conference (RE), 2017 IEEE 25th International*, pages 312–321. IEEE.

Charles J Fillmore et al. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32.

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson HS Liu, Matthew Peters, Michael Schmitz, and Luke S Zettlemoyer. 2017. A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2018. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Dan Jurafsky and James H Martin. 2014. *Speech and language processing*, volume 3. Pearson London.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Oge Marques and Nitish Barman. 2003. Semi-automatic semantic annotation of images using machine learning techniques. In *International semantic web conference*, pages 550–565. Springer.

Palmer Martha, Gildea Dan, and Kingsbury Paul. 2005. The proposition bank: a corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1).

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Helen Nissenbaum. 2009. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4949–4959, Hong Kong, China. Association for Computational Linguistics.

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

Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2014. Towards usable privacy policies: Semi-automatically extracting data practices from websites' privacy policies. *Poster Proceedings, SOUPS*, pages 9–11.

Madelyn Sanfilippo, Yan Shvartzshnaider, Irwin Reyes, Helen Nissenbaum, and Serge Egelman. 2019. Disaster privacy/privacy disaster. *Journal of the Association for Information Science and Technology (JASIST)*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. 2016. Automatic extraction of opt-out choices from privacy policies. In *AAAI Fall Symposium on Privacy and Language Technologies*.

Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2764–2769.

Peng Shi and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *CoRR*, abs/1904.05255.

Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2019a. Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 162–170.

Yan Shvartzshnaider, Zvonimir Pavlinovic, Ananth Balashankar, Thomas Wies, Lakshminarayanan Subramanian, Helen Nissenbaum, and Prateek Mittal. 2019b. Vaccine: Using contextual integrity for data leakage detection. In *The World Wide Web Conference*, pages 1702–1712.

Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning privacy expectations by crowdsourcing contextual informational norms. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Brett South, Shuying Shen, Jianwei Leng, Tyler Forbush, Scott DuVall, and Wendy Chapman. 2012. A prototype tool set to support machine-assisted annotation. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 130–139.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. 2022. OVRseen: Auditing network traffic and privacy policies in oculus VR. In *31st USENIX Security Symposium (USENIX Security 22)*, Boston, MA. USENIX Association.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016a. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1330–1340.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016b. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 133–143, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Martin Zurowietz, Daniel Langenkämper, Brett Hosking, Henry A Ruhl, and Tim W Nattkemper. 2018. Maia—a machine learning assisted image annotation method for environmental monitoring and exploration. *PloS one*, 13(11):e0207498.

Appendix

A Annotation reliability

We assessed the reliability of the expert by co-coding with another author annotating 10 of the 50 policies to achieve a substantial to excellent agreement with Cohen's Kappa score range of 0.67-0.88 on the CI parameter annotations. Table 7 shows the corresponding breakdown of Cohen's Kappa score (McHugh, 2012) for each of the CI parameters.

Parameter	**Cohen's Kappa Agreement score**
Sender	0.88
Receiver	0.80
Attribute	0.63
Subject	0.73
TP	0.68

Table 7: The Cohen's Kappa agreement scores for CI parameters annotation.

B Additional Tables

Parameter	**No. of Labels**
Sender	546
Subject	340
Attribute	1133
Receiver	928
TP	1101

Table 8: Annotated CI parameters labels

"Sending" action verbs	**"Receiving" action verbs**
share (1188)	acquire (27)
transfer (130)	learn (56)
exchange (7)	collect (1042)
show (49)	receive (376)
send (297)	use (346)
supply (59)	gather (78)
provide (1142)	include (629)
disclose(359)	contain (413)
submit (150)	ask (13@)
give (101)	accept (290)
deliver (109)	store (212)
mail (16)	require (504)
display (69)	save (30)
export (3)	record (118)
forward (22)	keep (77)
refer (60)	combine (51)
release (37)	get (21)
tell (44)	track (134)
pass (13)	

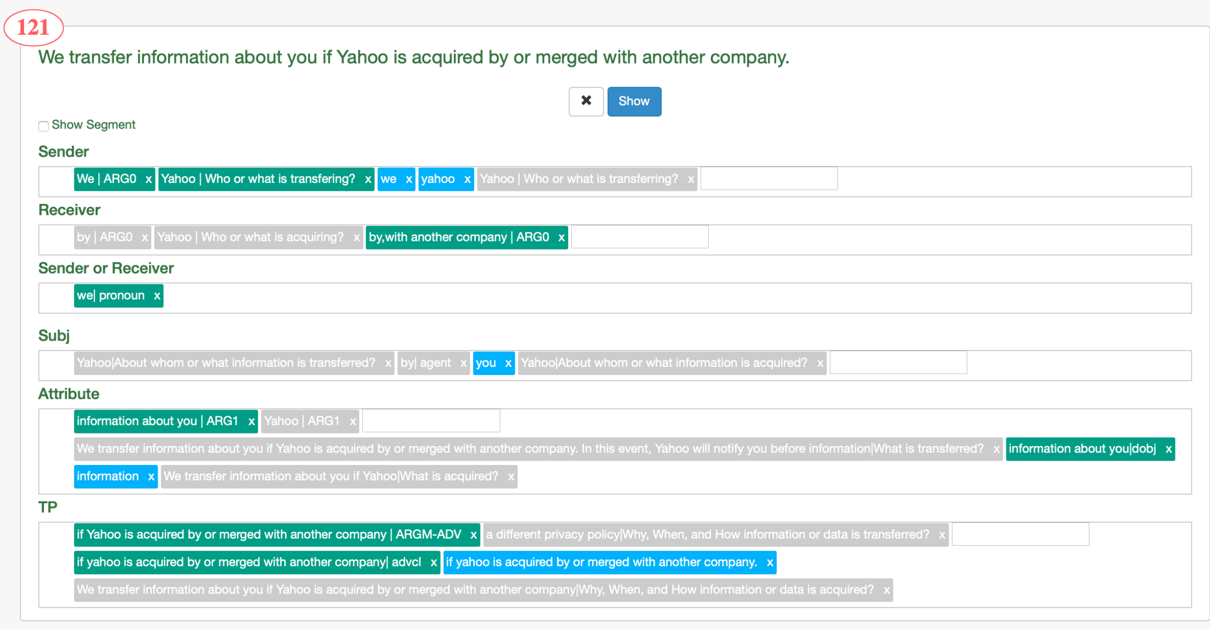
Table 11: Verbs used to trigger SRL semantic frames. The number in the brackets shows the frequency of appearance in the OPP-115 corpus.

Company	Recall	Precision	F1
Liquor	0.86	0.47	0.61
Lodgemfg	0.73	0.55	0.63
New Orleans Online	0.88	0.55	0.68
St. Louis fed	0.82	0.58	0.68
Google	0.79	0.67	0.73
UH	0.78	0.69	0.73
NBC Universal	0.80	0.67	0.73
Sheknows	0.87	0.67	0.75
Instagram	0.87	0.67	0.75
Gawker	0.87	0.67	0.76
Military	0.91	0.66	0.76
Citizen	0.87	0.68	0.76
Everydayhealth	0.84	0.72	0.77
FoxSports	0.94	0.66	0.78
CoffeeReview	0.88	0.70	0.78
Reference	0.95	0.67	0.79
Zacks	0.93	0.69	0.79
News Busters	0.87	0.73	0.79
DCCCD	0.81	0.78	0.80
PlayStation	0.88	0.73	0.80
ABCNews	0.88	0.74	0.80
Yahoo	0.81	0.80	0.80
High Gear Media	0.92	0.72	0.81
PBS	0.85	0.77	0.81
Fortune	0.93	0.72	0.81
Time inc	0.94	0.73	0.82
LatinPost	0.90	0.77	0.83
Lynda	0.90	0.78	0.84
IMDB	0.88	0.80	0.84
MSN	0.87	0.81	0.84
Reddit	0.89	0.80	0.84
Ted	0.97	0.77	0.86
NY Times	0.87	0.86	0.87
AOL	0.98	0.78	0.87
Geocaching	0.91	0.82	0.87
Amazon	0.90	0.84	0.87
TaylorSwift	0.96	0.81	0.88
USA	0.96	0.81	0.88
WashingtonPost	0.94	0.83	0.88
TheAtlantic	0.96	0.83	0.89
Dogbreedinfo	0.92	0.86	0.89
Austincc	0.93	0.85	0.89
Walmart	0.93	0.85	0.89
BankofAmerica	0.87	0.92	0.90
Fool	0.95	0.85	0.90
TicketMaster	0.91	0.89	0.90
SI	0.90	0.93	0.91
TheFreeDictionary	0.92	0.93	0.92
Earthkam	1.00	0.88	0.93
OpenSecrets	1.00	1.00	1.00

Table 9: F1 scores for all annotated policies

Model	CI Parameters	Recall	Precision	F1	TP	FP	FN
DP	Attribute	0.35	0.71	0.47	791	1490	322
	Subject	0.24	0.75	0.36	255	806	85
	TP	0.57	0.81	0.67	895	666	206
SRL (0.8.5)	Attribute	0.55	0.85	0.67	960	773	173
	Receiver	0.67	0.85	0.75	790	394	138
	Sender	0.55	0.86	0.67	470	381	76
	TP	0.62	0.74	0.67	814	498	287
SRL (2.1)	Attribute	0.64	0.66	0.74	952	626	181
	Receiver	0.70	0.85	0.74	787	335	141
	Sender	0.70	0.85	0.77	460	288	86
	TP	0.72	0.78	0.75	855	325	246
CI-SRL (0.8.5)	Attribute	0.60	0.83	0.7	939	617	194
	Receiver	0.72	0.81	0.76	754	298	174
	Sender	0.64	0.83	0.72	453	251	93
	TP	0.76	0.73	0.74	805	258	296
CI-SRL (2.1)	Attribute	0.69	0.81	0.75	923	416	210
	Receiver	0.78	0.80	0.79	743	212	185
	Sender	0.74	0.79	0.77	433	149	113
	TP	0.85	0.75	0.79	821	144	280

Table 10: F1 Scores for all the models: DP, SRL models and Improved SRL (CI-SRL). Here, the recall and precision values for each parameter is calculated using the total number of parameters, and not macro averaging over privacy statements as in Table 7. TP, FP and FN are the number of True Positive (TP) and False Positive (FP).



121

We transfer information about you if Yahoo is acquired by or merged with another company.

Sender

We | ARG0 x · Yahoo | Who or what is transfering? x · we x · yahoo x · Yahoo | Who or what is transferring? x

Receiver

by | ARG0 x · Yahoo | Who or what is acquiring? x · by,with another company | ARG0 x

Sender or Receiver

we| pronoun x

Subj

Yahoo|About whom or what information is transferred? x · by| agent x · you x · Yahoo|About whom or what information is acquired? x

Attribute

information about you | ARG1 x · Yahoo | ARG1 x · We transfer information about you if Yahoo is acquired by or merged with another company. In this event, Yahoo will notify you before information|What is transferred? x · information about you|dobj x · information x · We transfer information about you if Yahoo|What is acquired? x

TP

if Yahoo is acquired by or merged with another company | ARGM-ADV x · a different privacy policy|Why, When, and How information or data is transferred? x · if yahoo is acquired by or merged with another company| advcl x · if yahoo is acquired by or merged with another company. x · We transfer information about you if Yahoo is acquired by or merged with another company|Why, When, and How information or data is acquired? x

Figure 3: Annotation interface. The expert annotator marks which outputs matched the expectation and criteria for the five CI parameters.