

# RECIPE: Applying Open Domain Question Answering to Privacy Policies

**Yan Shvartzshnaider**

New York University and Princeton University  
yansh@nyu.edu

**Ananth Balashankar**

New York University  
ananth@nyu.edu

**Thomas Wies**

New York University  
wies@cs.nyu.edu

**Lakshminarayanan Subramanian**

New York University  
lakshmi@cs.nyu.edu

## Abstract

We describe our experiences in using an open domain question answering model (Chen et al., 2017) to evaluate an out-of-domain QA task of assisting in analyzing privacy policies of companies. Specifically, Relevant CI Parameters Extractor (RECIPE) seeks to answer questions posed by the theory of contextual integrity (CI) regarding the information flows described in the privacy statements. These questions have a simple syntactic structure and the answers are factoids or descriptive in nature. The model achieved an F1 score of 72.33, but we noticed that combining the results of this model with a neural dependency parser based approach yields a significantly higher F1 score of 92.35 compared to manual annotations. This indicates that future work which incorporates signals from parsing like NLP tasks more explicitly can generalize better on out-of-domain tasks.

## 1 Introduction

Open domain question answering approaches offer a promising glimpse into a future in which machines are able to perform sophisticated cognitive tasks on behalf of a human. Recent advances in deep neural networks applied to reading comprehension and document retrieval (Chen et al., 2017; Wang et al., 2017) have achieved competitive results for answering questions based on a large and diverse corpus of documents. These models are trained on Wikipedia text and are capable of answering various factoid and descriptive questions. Specifically, the distribution of the interrogative words used, lexical and syntactic variations, reasoning across multiple sentences and ambiguous

statements in such Wikipedia datasets (Ryu et al., 2014; Rajpurkar et al., 2016) results in a robust QA model.

Motivated by their success, we set to apply one such model (Chen et al., 2017) to evaluate an out-of-domain QA task to assist in analysis and understanding of privacy policies. The text in privacy policies is notoriously cumbersome, confusing and hard to comprehend even for legal experts (Reidenberg et al., 2015). So when it comes to reading privacy policies, users often miss important information or skip reading them altogether. Past efforts have applied NLP (Sathyendra et al., 2017, 2016; Evans et al., 2017), Machine Learning (Ammar et al., 2012) and crowdsourcing (Wilson et al., 2016b) techniques to identify important and relevant privacy statements. However, identifying paragraphs that mention sensitive information in privacy policies only takes us half-way because we still need to understand who collects that information, who receives it, and under what conditions the collection happens. To support this finer-grained analysis, we present a case for using a machine comprehension model for answering questions posed by the theory of contextual integrity (CI) (Nissenbaum, 2010).

To understand the privacy implications, the CI theory (see Section 2) calls for answering to the following questions: who are the actors (sender, recipient, subject) involved in the information flow? What are the type of information (attribute), and condition (transmission principle) under which information is exchanged? The answers are used to formulate information flows (CI flows) into a five element tuple (sender, attribute, receiver, subject, transmission principle).

Unfortunately, because privacy policies are not written with CI in mind, manually identifying CI flows in text is a time consuming exercise. It requires substantial cognitive effort to understand

and answer CI related questions. Automating this task is not trivial either. The syntax of privacy statements varies a lot, some relevant information might not be specified at all. So, the answers are not always obvious and cannot be identified using a simple model. Even with the simple syntactic structure of the questions, the trained model needs to take into account the variations and complexity in syntax and semantics used in privacy policies.

To help address these issues, we are designing a Relevant CI Parameters Extractor (RECIPE) that uses a pre-trained 3 layer bi-directional LSTM reading comprehension (RC) model (Chen et al., 2017). Our experiments show that, in itself, the model achieved an F1 score of 72.33 for identifying answers in the text. However, the results significantly improved by combining the results from the model and that of a neural dependency parser. The combination of the two approaches yielded an overall F1 score of 92.35 against the baseline of six manually annotated privacy policies.

## 2 Contextual Integrity Primer

Questions underpinning the theory of Contextual Integrity (CI) are used by many research efforts for understanding privacy implications in a given context. Legal and privacy scholars draw on CI to examine existing data sharing practices in companies like Facebook (Hull et al., 2011) and Google (Zimmer, 2008) and to identify important contextual elements behind users privacy expectations (Martin and Nissenbaum, 2016). In computer science, CI has been used to build privacy compliance and verification tools (Barth et al., 2006; Chowdhury et al., 2013).

The CI theory defines privacy as appropriate flow of information. The appropriateness is determined by established norms in a given context. CI offers a framework to capture the information flow and contextual norms using a five element tuple that specifies the following CI parameters: (*sender, attribute, receiver, subject, transmission principle*). Therefore, answering the questions about who are the actors (sender, recipient, subject) involved in the information flow, what are the type of information (attribute), and condition (transmission principle) of the information exchange, is crucial for identifying potential privacy violations.

In our work, the answers to the questions are highlighted in the text, as shown in Figure 1 to

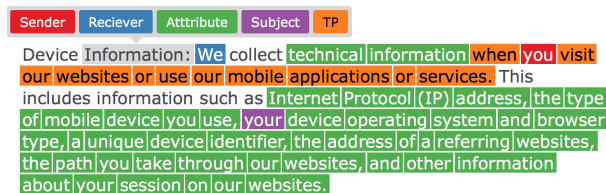


Figure 1: Example of a Walmart privacy statement. The coloring of the text identifies the CI parameters.

assist consumers in understanding the information flows described in the privacy policy. These annotations can also help identify potentially confusing or misleading statements, e.g., cases where one of the five parameters such as transmission principle or receiver is missing or ambiguous.

## 3 Related work

The problem of automatic parsing of privacy policies was tackled by multiple efforts, as part of the “Usable Privacy Project” (Sadeh et al., 2013), which included training machine learning models to identify paragraphs containing specific data practices information (Wilson et al., 2016b), identifying provision of choice statements, automatic extraction of the opt-out choice statements (Sathyendra et al., 2016, 2017) and others (Evans et al., 2017; Hosseini et al., 2016; Bhatia et al., 2016b). More recent work by Harkous et al. (2018) looked at neural network classifiers for annotation and support of free-form querying of the privacy policy content.

Compared to these efforts, our work is the first step towards bringing a formal analysis of privacy policies through the use of CI. We used dependency parsing to extract CI parameters from individual sentences, which was previously successfully used for various privacy policy analysis tasks. Bhatia and Breau (2015) used dependency types to construct an information type lexicon from manual human annotations and an entity extractor based on part-of-speech tagging. In a more recent work, Bhatia et al. (2016a) employed typed dependency parsing to extract privacy goals from privacy policies. In our work, we were able to show that a machine reading comprehension model (Chen et al., 2017; Hermann et al., 2015; Seo et al., 2017), needs to additionally have the capacity to capture the semantic relationships as-

sociated with dependency parsing structure from the privacy policy text.

Furthermore, the problem of identifying CI parameters can be modeled as a sequence tagging task using LSTMs combined with CRFs (Lample et al., 2016; Ma and Hovy, 2016) to learn classification models for entity recognition and end-to-end sequence labeling. However, training these models requires large amounts of annotated data, which in our case means hiring highly skilled (i.e., expensive) annotators to produce a high quality corpus of CI-based annotated policies. In fact, our work aims at reducing that effort.

#### 4 Relevant CI Parameters Extractor (RECIPE) using open-domain QA

In this section we describe our approach of formulating the CI parameter extraction as an open-domain QA task. Given a privacy statement paragraph, such as shown in Figure 1, we pose questions that correspond to each of the CI parameters, i.e., what type of information being exchange (“technical information”), who is the sender (“You”), receiver (“We”) and subject (“your”) involved in the information exchange and under what conditions (“when visiting websites or using a mobile application service”)?

As in the example paragraph, some parameters can be mentioned elsewhere, e.g., in the next sentence. To identify CI parameters at a paragraph level using global relationships across sentences, we used a 3 layer bi-directional LSTM reading comprehension (RC) model (Chen et al., 2017). The model uses features of Glove word embeddings trained from 840B Web crawl data (Pennington et al., 2014) and token features like presence of exact match with question’s tokens, parts of speech, named entity relationship and term frequency. We also align the question embedding with the paragraph token embedding using soft attention (Lee et al., 2016). During the prediction phase, we perform two classification tasks to predict the beginning and ending of the span of the answer. The model is trained on the SQuAD dataset (Rajpurkar et al., 2016) for machine comprehension (87k examples for training and 10k for development) based on Wikipedia. Each question in the dataset has an answer which is contained within the paragraph. In our case, the answer span detection is identical to the task of manually annotating CI parameters. In order to extract the 5

Parameter Type	Question
Attribute	What is the information?
Receiver	Who is receiving the information?
Sender	Who is sending the information?
Transmission Principle	When is the information sent?
Subject	Whose information is it?

Table 1: Questions asked for each contextual integrity parameter

parameters, we ask a fixed set of questions for a given paragraph as can be seen in Table 1. The top 5 answers from the reading comprehension model are then used to evaluate the accuracy via fuzzy string token match above a certain threshold (>80%). We again manually validate if the answers provided by the model match the manually annotated answers semantically. For example during extraction of “receiver” in the New York Times policy, the model answered with the entity “New York Times” whereas it was manually annotated as “we”, which referred to NYT in the context. This manual validation is required due to the ambiguity prevalent in the privacy policy text and relying on exact match can be misleading during evaluation.

In order to understand the limitations of this model’s performance (discussed in Section 5), we compared the results produced by dependency parsing which is limited to identifying CI parameters at a single sentence level using local relationships. We discuss the dependency-parsing approach next.

---

#### Algorithm 1

---

```

paras ← CI relevant paragraphs
sentences ← segments.split('.')
CIDP ← DependencyParser(sentences)
CIRC ← ReadingCompr(paras)
Return CIDP ∪ CIRC

```

---

#### 4.1 Dependency parsing

We run a typed dependency parser (DP) on the text of the policies by splitting the paragraph into individual sentences and then parsing each sentence using the Spacy I/O<sup>1</sup> dependency parser. The library (Honnibal and Montani, 2018) achieves near state of the art performance on most NLP tasks.<sup>2</sup> We then match dependency types to specific CI parameters. For example, in the sentence “You

<sup>1</sup><https://spacy.io>

<sup>2</sup><https://spacy.io/usage/facts-figures>

may be asked to provide your personal information anytime you are in contact with Apple or an Apple affiliated company: The nominal subject “you” and conjunct “Apple affiliated company” are sender and receiver respectively, the direct object “your personal information” is an attribute, and the open clausal complement adverbial clause modifier “anytime you are in contact with Apple or an Apple affiliated company” is the transmission principle.

Table 2 shows all the mappings of dependency types to CI parameters. For the sake of brevity, we represent type dependencies as they are defined in the Stanford Typed Dependency Manual (De Marneffe and Manning). We then present the options of the CI parameters as extracted by the dependency parsers to the annotators, who then validate them as explained in the previous section, without having an option to modify the text for the annotation.

CI Parameter Type	Dependency types
Attribute	<i>doj, parataxis, nsubjpass</i>
Sender/Receiver	<i>nsubj, pronouns</i>
Transmission Principle	<i>xcomp, ccomp, advcl, oprd</i>
Subject	<i>poss, agent</i>

Table 2: Mapping of dependency types corresponding to CI parameters

In summary, in our approach we rely on dependency parsing to extract CI parameters based on the syntactic structure of a single sentence, and a reading comprehension model to capture CI parameters based on the semantic understanding of the larger scope. More specifically, the approach consists of two rounds of extraction, as shown in Algorithm 1, to return the union of the two sets of parameters after manual validation assisted by a fuzzy token match (Cohen, 2011) that uses Levenshtein Distance to calculate the differences between sequences.

## 5 Evaluation

For our evaluation, we analyze the ensemble algorithm against each of its sub-algorithms using manually annotated policies as a baseline.

### 5.1 Dataset

We use OPP-115 Corpus (Wilson et al., 2016a) that contains website privacy policies in natural text along side annotations (done by law students) specifying the corresponding text data prac-

Parameter Type	F1 score	Validated F1 score
Attribute	61.53	83.65
Receiver	48.57	75.00
Sender	37.70	56.55
Transmission Principle	52.20	67.29
Subject	25.00	26.56
Overall	49.20	72.33

Table 3: Validated F1 score of manually validated reading comprehension based annotation

Parameter Type	Valid (%)	Validated F1 score
Attribute	23.73	84.39
Sender/Receiver	40.41	73.04
Subject	17.42	86.05
Transmission Principle	27.47	69.48
Overall	26.40	77.07

Table 4: F1 scores of dependency parsing based annotation

tices. We rely on these labels to extract segments within the privacy statement about information exchanges. Specifically, we chose segments related to First Party Collection/Use, Third party sharing/collection, Data Retention to extract CI parameters. We then manually annotated a total of 715 parameters (219 Attributes, 65 Subjects, 164 Transmission Principles, 124 Senders and 143 Receivers) from 6 privacy policies from Amazon, Google, New York Times, The Atlantic, Bank of America and Walmart. This forms the ground truth of our evaluation.

## 5.2 Results and Discussion

Table 3 shows the F1 scores of the reading comprehension model alone. We see that subject, receiver and sender have relatively low F1 scores and this is due to the absence of entities in the paragraphs as they are usually referred to as pronouns. We noticed that the reading comprehension model outputs such pronoun answers only for 15% of these parameters and prefers to answer with often incorrect entities for these questions.

Table 4 shows the results from using the dependency parser on its own. Since dependency parsing is restricted to a single sentence, it misses out on some contextual information spread across sentences as shown in Figure 1, leading to a loss in accuracy, even after manual validation.

Table 5 compares the results from each algorithm individually with the result achieved by ensembling the approaches. Exclusive contributions

Parameter	DP only	RC only	DP $\cap$ RC	DP $\cup$ RC
Attribute	12.98	16.34	67.31	96.63
Sender/Receiver	29.57	30.28	33.1	92.95
Transmission Principle	20.75	23.27	44.02	88.05
Subject	57.14	19.05	7.14	83.33
Overall	24.24	23.81	44.30	92.35

Table 5: Validated F1 score by ensembling

by each of the individual models (“Parsing only”, “RC only”) demonstrate how each model is able to capture syntactic and semantic features respectively, which the other could not. Column “DP  $\cap$  RC” shows that many parameters are extracted by both the approaches used to formulate CI parameter extraction. This confirms that the task of CI parameter extraction is non-trivial and composes semantic and syntactic relationship extraction within it. Finally, “DP  $\cup$  RC” achieves the highest F1 score, significantly improving over both of the component scores individually.

Table 6 shows the number of incorrect answers yielded by the hybrid approach, i.e., the answers where neither RC nor DP performed well.

Parameter	Incorrect (%)
Sender	4 (7.55 %)
Subject	7 (13.21 %)
Attribute	7 (13.21 %)
Receiver	16 (30.19 %)
Transmission Principle	19 (35.85 %)
Overall	53

Table 6: A summary of inaccurate labeling by the ensembling approach. The percentage is out of total number of invalid labels.

The inaccurate results for these answers can be partially attributed to the fact that our approach has to produce an answer (label) each time, which in some cases is not explicitly stated in the text. Furthermore, some paragraphs have a complex syntactic structure, comprising very long sentences glued together by semicolons. This poses a challenge to both DP and RC, which rely on the sentence structure and model trained on syntactically proper sentences, respectively, to identify the relevant parameters. Dealing with such semantic and syntactic ambiguities while correctly identifying relevant entities remains an open research question.

## 6 Conclusion

In this paper we present our work towards designing a Relevant CI Parameters Extractor (RECIPE) that leverages an open domain QA model on Privacy Policies to answer questions posed by CI. The theory of CI relies on identifying five parameters to reason about privacy violation. This is done by asking questions about who are the actors (sender, recipient, subject), what type of information being conveyed and under what conditions in an information flow.

This paper evaluates an open-domain QA model to find answers to contextual integrity questions in privacy policy texts, a complex task which is traditionally delegated to legal and privacy experts. We perform an error analysis of the model on this out-of-domain task and show that this error can be reduced by combining the open domain QA model with a neural dependency parser.

Through this paper, we hope to motivate the community to incorporate signals from dependency parsing like NLP tasks more explicitly in solving complex tasks like open-domain QA for generalizing on out-of-domain tasks.

## 7 Acknowledgments

We thank Schrasing Tong for his help in the initial stage of this work. This work is supported by the National Science Foundation under grant CCF-1350574.

## References

- Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic Categorization of Privacy Policies: A Pilot Study. *Carnegie Mellon University Technical Report*, (CMU-ISR-12-114).
- Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. 2006. Privacy and Contextual Integrity: Framework and Applications. In *IEEE Symposium on Security and Privacy*, pages 184–198.
- Jaspreet Bhatia and Travis D Breaux. 2015. Towards an Information Type Lexicon for Privacy Policies. In *IEEE Eighth International Workshop on Requirements Engineering and Law (RELAW)*, pages 19–24.
- Jaspreet Bhatia, Travis D Breaux, and Florian Schaub. 2016a. Mining Privacy Goals from Privacy Policies Using Hybridized Task Recomposition. In *ACM Transactions on Software Engineering and Methodology (TOSEM)*, pages 22:1–22:24.
- Jaspreet Bhatia, Morgan C Evans, Sudarshan Wadkar, and Travis D Breaux. 2016b. Automated Extraction

- of Regulated Information Types using Hyponymy Relations. In *IEEE International Requirements Engineering Conference Workshops (REW)*, pages 19–25.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Association for Computational Linguistics (ACL)*.
- Omar Chowdhury, Andreas Gampe, Jianwei Niu, Jeffrey von Ronne, Jared Bennett, Anupam Datta, Limin Jia, and William H Winsborough. 2013. Privacy Promises That Can Be Kept: A Policy Analysis Method with Application to the HIPAA Privacy Rule. In *Proceedings of the 18th ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 3–14.
- A Cohen. 2011. FuzzyWuzzy: Fuzzy string matching in python. <https://github.com/seatgeek/fuzzywuzzy>.
- Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. [https://nlp.stanford.edu/software/dependencies\\_manual.pdf](https://nlp.stanford.edu/software/dependencies_manual.pdf).
- Morgan C Evans, Jaspreet Bhatia, Sudarshan Wadkar, and Travis D Breau. 2017. An Evaluation of Constituency-based Hyponymy Extraction from Privacy Policies. In *IEEE 25<sup>th</sup> International Requirements Engineering Conference (RE)*, pages 312–321.
- Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. *arXiv preprint arXiv:1802.02561*.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1693–1701.
- Matthew Honnibal and Ines Montani. 2018. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. <https://spacy.io>.
- Mitra Bokaei Hosseini, Sudarshan Wadkar, Travis D Breau, and Jianwei Niu. 2016. Lexical Similarity of Information Type Hypernyms, Meronyms and Synonyms in Privacy Policies. In *AAAI Fall Symposium Series*.
- Gordon Hull, Heather Richter Lipford, and Celine Latulipe. 2011. Contextual gaps: privacy issues on facebook. *Ethics and Information Technology*, 13:289–302.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. *Neural Architectures for Named Entity Recognition*. *CoRR*, abs/1603.01360.
- Kenton Lee, Tom Kwiatkowski, Ankur P. Parikh, and Dipanjan Das. 2016. Learning recurrent span representations for extractive question answering. *CoRR*, abs/1611.01436.
- Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.
- Kirsten Martin and Helen Nissenbaum. 2016. Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables. *Columbia Science and Technology Law Review*, 18:176.
- Helen Nissenbaum. 2010. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Joel R Reidenberg, Travis Breau, 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.
- Pum-Mo Ryu, Myung-Gil Jang, and Hyun-Ki Kim. 2014. Open domain question answering using Wikipedia-based knowledge model. *Information Processing & Management*, 50(5):683–692.
- Norman Sadeh, Alessandro Acquisti, Travis D Breau, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The Usable Privacy Policy Project. *Technical report, CMU-ISR-13-119, Carnegie Mellon University*.
- 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 *Conference on Empirical Methods in Natural Language Processing*, pages 2764–2769.

- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *5<sup>th</sup> International Conference on Learning Representations (ICLR)*.
- Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2017. [R<sup>3</sup>: Reinforced Reader-Ranker for Open-Domain Question Answering](#). *CoRR*, abs/1709.00023.
- 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 *54<sup>th</sup> 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 *25<sup>th</sup> International Conference on World Wide Web*, (WWW), pages 133–143.
- Michael Zimmer. 2008. Privacy on Planet Google: Using the Theory of “Contextual Integrity” to Clarify the Privacy Threats of Google’s Quest for the Perfect Search Engine. *Journal of Business Technology Law*, 3:109.