Automatic Section Title Generation to Improve the Readability of Privacy Policies

Abhijith Athreya Mysore Gopinath
Pennsylvania State University

Vinayshekhar Bannihatti Kumar
Carnegie Mellon University

Shomir Wilson
Pennsylvania State University

Norman Sadeh
Carnegie Mellon University

Abstract

The text in a typical privacy policy is organized into different topics, either implicitly or explicitly. Few policies with explicit visual organization split the text into various sections with titles that briefly describe the paragraphs that follow. Prior studies prove that the explicit visual organization improves the readability of privacy policies. However, policies with implicit organization either lack visual cues or do not contain titles for paragraphs, rendering them difficult to read. To make these policies more readable, we work on the automatic generation of titles for paragraphs. We exploit the explicit structure of web privacy policies to generate a title-paragraph dataset and train supervised generative models that generate a title for a given chunk of the privacy policy text. We score the models on both automatic and human evaluation metrics and observe that the Transformer based approach outperforms sequential models producing relevant and grammatical outputs for 48% of the test cases.

1 Introduction

Privacy policies are essential documents that contain vital information regarding the collection, sharing, and usage of customer data, but internet users face difficulties in understanding them [18]. Many users do not bother to read them [6, 13]. Reading becomes arduous because of the lengthy nature of these documents, and the use of legal jargon, which is tough to comprehend. There is a need to bring privacy policies into some intermediate representation, which the users can easily understand. This problem provides an opportunity to use and develop natural language processing, machine learning, and information retrieval methods to lessen the divide between privacy policies and end-users.

A typical privacy policy contains information related to various aspects of the company's privacy terms. These aspects can relate to different data practices and company-specific topics. In some privacy policies, topics are manifested as titles, containing a short text which holds a summary for the following paragraph of text. Automatic methods exist that extract titles from these kinds of policies. However, textual privacy policies do not contain any explicit organization of title-paragraph structures and require the dynamic generation of titles. Traditional topic detection is less expressive, requires labeled data that contains sentence (or paragraphs)-topic pairs. The labeling of privacy policy data is expensive as reliable results typically involve the services of legal experts in place of unspecialized crowd workers. Dynamic title generation, on the other hand, is specific, helps in the generation of better titles that reflect the changing privacy landscape and emerging concerns. For example, Figure 1 contains a general topic and a generated title for the text[1] present in the inner rectangle. The generated title contains more information, is more intuitive and semantically closer to the paragraph than the generalized topic.

Title generation can help in the creation of a navigable table of contents section for privacy policies, thus improving its readability. On the computational side, it leads to several advantages for NLP tasks on privacy policies. It can reduce the search space in information retrieval tasks, improving the system's speed and accuracy. It can also aid question answering on privacy policies by identifying maximally relevant sections to search for answers. Instead of looking through the entire document, a high-level search can be performed over the titles and a deep search over the actual text. The close semantic similarity of titles with underlying paragraphs makes them better candidates for search and retrieval tasks.

Title generation poses significant challenges that sepa-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2020.
August 9–11, 2020, Boston, MA, USA.

[1] Obtained from https://about.9gag.com/privacy/

Topic: First Party Collection

Generated Title: *Information we get when you use our services*

When you use our services, we collect information about which services you use and how you use them. For example, we may know what section or post you are reading, when you're reading it, and how you're reading it (desktop, mobile browser, or mobile app).

Figure 1: Topic and generated title for the text contained in the inner box.

rate it from a straightforward sequence to sequence learning task. First, there is no readily available data in the form of paragraph-title pairs to train a supervised model that learns to generate titles for a given paragraph. Second, the text in privacy policies is more cohesive than open-domain text as it pertains to specificities of privacy in general, resulting in similar sentences. The presence of similar sentences across various sections of the document results in minimal computational differences between them, and the generation of diverse titles for similar but different paragraphs becomes difficult.

We overcome the challenges of title generation and present a Transformer [26] based model that generates a title for a given privacy policy text. We make use of ASDUS [11], a system that separates section titles from prose text in web documents to generate paragraph-title pairs on a corpus of 150 privacy policies. We then use these paragraph-title pairs to train a Transformer model with domain-specific embeddings that learns to generate micro summaries of the privacy policy sentences. In automatic evaluation, our method achieves a ROUGE score of 35.96 and a semantic similarity score of 47.31. In human evaluation, it scores 63% in fidelity and 90% in fluency. Our system effectively generates titles for each paragraph (or sentence) of a given privacy policy. The generated titles provide a chance to quickly glance over the contents of the policy and act as location markers for specific paragraphs, thus enhancing the reading experience and usability of privacy policies.

2 Related Work

No prior work was found to exist on section title generation for privacy policies, although some similar work exists on topic detection and headline generation, as we describe below.

2.1 Topic Detection in Privacy Policies

Wilson and others created the OPP-115 corpus that contains manual annotations for numerous data practices [27]. They identified the ten most commonly found data practices (loosely, *topics*) and divided them further into category-specific attributes. Expert annotators labeled the text spans of 115 privacy policies as belonging to different data practices and attributes. Liu and others use the OPP-115 corpus to build a supervised model that classifies each segment of a privacy policy into different data practices [16]. They also build unsupervised topic models using Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). Choi and others perform topic modeling on privacy related documents along different dimensions [4]. They compare and contrast the vocabulary usage over different business sectors at different points of time. Through this study, they identify 47 popular topics in privacy related documents.

Sarne and others perform unsupervised topic extraction over 4,982 privacy policies downloaded from the Google Play Store [20]. They treat each paragraph as a separate document and apply LDA at different thresholds. Then they make use of an expert annotator to merge the 82 automatically extracted topics into 36 different topics. They map these 36 topics to the ten data practices identified by Wilson and others in the creation of OPP-115 corpus. Liu and others create the APPCorp, an annotated corpus consisting of 167 privacy policies [17]. They manually annotate the paragraphs and sentences of all privacy policies with one of the 11 labels. They also use neural models to perform automatic topic classification on the labeled data and achieve greater than 80% accuracy with a BERT [7] classification model.

2.2 Headline Generation

Headline generation is similar to abstractive text summarization with the difference in the length of the output. In most cases, headlines are much shorter in length than summaries.

Text generation techniques make use of an encoder-decoder architecture, wherein the encoder computes a representation of the input, and the decoder generates the output. Prior work employs sequence to sequence based Long Short Term Memory (LSTM) celled encoder-decoder architecture to generate headlines for news articles [12, 23]. Traditional sequence to sequence models fail to capture long range dependencies resulting in poor headlines for long articles. To remedy this, attention layers are added to the encoder-decoder networks, thus resulting in better headlines [28]. Even with the addition of layers, sequence based models take longer to train due to the serial nature of execution. The Transformer [26] allows parallel training with self-attention mechanism enabling faster training and learning of complex dependencies. Transformer based headline generators fare better than sequence to sequence models [9].

Low resource datasets lack a substantial amount of examples to train large generative networks. Hence, the decoder suffers from limited vocabulary exposure producing similar outputs. One solution is to pre-train a decoder on a bigger

generic corpus, thus exposing it to a large vocabulary set [24]. Datasets with large paragraphs are difficult to train, and sentence ranking methods are used to reduce the size, resulting in shorter yet useful training data [10].

3 Dataset

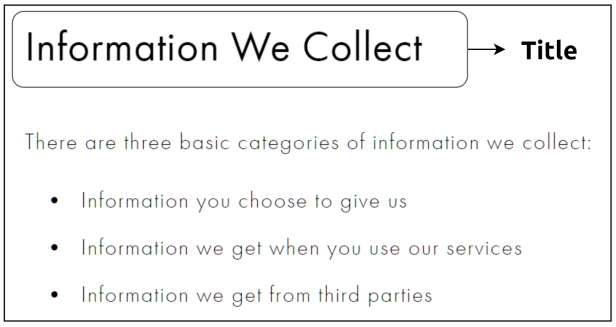

Figure 2: A partial screengrab of a web privacy policy indicating the title followed by a paragraph.

We need sentence-title and paragraph-title pairs to train a supervised model for title generation. However, there are no readily available datasets in this form. We make use of the internal structure of web privacy policies to create a suitable dataset. The textual content of most online privacy policies contains a title or heading that is visually different from the following paragraph text. Figure 2 contains a portion of a web privacy policy[2] depicting the visual separation of headings from paragraphs. Automatic collection of these headings and paragraphs is not straightforward as the HTML structure of web privacy policies differs from one another. Therefore, we make use of ASDUS [11] to segregate headings and paragraphs of 152 privacy policies and prepare the paragraph-title dataset. We perform sentence segmentation on the individual paragraphs to prepare the sentence-title dataset. The paragraph-title dataset contains 3,504 instances, and the sentence-title dataset contains 25,800 instances. We use an 80:20 split for training and evaluation, respectively.

4 Approach

We formulate the task of title generation along the lines of machine translation. Our goal is to generate a contextually relevant headline (title) for the given privacy policy text.

4.1 Baseline Models

We define two deep learning based baseline models to compare our results. The first is a neural sequence to sequence model [22] that utilizes Gated Recurrent Units [3] in both the encoder and decoder. The encoder converts the input sequence into a *context vector*, and the decoder uses this context vector to output the target sequence one word at a time. The second baseline model is similar to the first, but with an extra attention layer [1] added to the encoder and decoder parts.

4.2 Transformer Model

We employ the Transformer [26] architecture as our main approach for title generation. We tokenize the input using a subword encoder to decrease the out of vocabulary (OOV) instances. The Transformer consists of an encoder, a decoder, and a final dense layer for making predictions. Both the encoder and decoder utilizes positional encoding to attend to different positions of words in the sentence. The input for the encoder is the subword embedding of the paragraph, and it computes an intermediate representation of the input using N identical layers stacked over one another. Each layer of the encoder consists of a multi-head attention layer and a pointwise feed forward network with residual connections around them to avoid vanishing gradients. Each of the N decoder layers consists of a multi-head attention layer with a look ahead mask, multi-head attention, and pointwise feed forward networks. The input to the decoder is the subword embedding of the output (topic). The decoder also receives the encoder's output and learns to predict the next word from attending to its input and the encoder's output.

We use four encoder and decoder layers, each containing four attention heads. We used 128-dimensional inputs and 512 units in feed forward layers with a dropout of 0.1. We used a custom learning rate and an Adam [14] optimizer.

5 Results

We present the evaluation criteria, followed by the results.

5.1 Evaluation Metrics

Automatic Evaluation: We use the standard machine translation metrics, the cumulative 4-gram BLEU [19] score, and the ROUGE [15] score to evaluate the system. BLEU and ROUGE measure the direct word to word similarity between the generated title and the reference title. In an ideal setting, these scores are derived from computing n-gram overlap of the generated text with several reference candidates to account for the variation in vocabulary. In our case, the presence of a single reference title limits the usefulness of these scores but provides a method to compare different models. Since these scores rely on word overlap metrics, they cannot accommodate variations in single candidate datasets [8, 21, 25]. A paragraph can have multiple right titles that are closer in meaning but consisting of different words. To accommodate this, we use the semantic similarity metric, which calculates

[2] Obtained from https://about.9gag.com/privacy/

Model	BLEU	ROUGE-1	Semantic Similarity	Human	
				Fidelity	Fluency
Baseline 1: Seq2Seq	6.78	34.61	30.32	39.00	61.00
Baseline 2: Seq2Seq+Attention	9.54	35.60	35.60	58.00	87.00
Transformer-Sentence	**13.25**	**35.96**	**47.31**	**63.00**	**90.00**
Baseline 2: Seq2Seq+Attention-Paragraph	7.27	26.32	24.72	19.00	67.00
Transformer-Paragraph	**9.55**	**28.01**	**36.31**	**48.00**	**91.00**

Table 1: Results of the automatic and human evaluation of all the models. Higher values indicate better performances.

the semantic distance between the generated title and the reference title using the universal sentence encoder [2].

Human Evaluation: Automatic evaluation does not entirely account for the grammar and diversity of the output. We perform a human evaluation of 100 random outputs to understand the system's functional performance. Two graduate students score each output on two criteria: fidelity and fluency. Fidelity depicts the relevance of the output, and fluency represents the grammatical correctness. We score both the metrics on a binary scale with zero being the least, and one being the highest. The Cohen's kappa score [5] for all the human evaluation metrics was 0.81, indicating substantial agreement.

5.2 Results

Table 1 contains the results of all the models with the first three rows belonging to the sentence-title pairs and the final two rows to the paragraph-title set. The low BLEU scores are due to the availability of only one reference title for each data instance. There is a negligible difference in the ROUGE-1 scores between different models for the same dataset. This might indicate that the models' output quality is similar, but that is not the case. In abstractive summarization tasks, there are multiple correct answers, and comparing each candidate answer to one reference is shortsighted. Consider the reference title *data security* and the candidate reference *information protection*, even though both are closely related to each other, the BLEU and ROUGE metrics arrive at a zero score for this pair, whereas the semantic similarity of this pair is 47. The Transformer model achieves greater than 30% higher semantic similarity scores than the baseline models in both the dataset variants. This difference is even more prominent in the paragraph dataset as the large size of the input brings out the inability of sequence to sequence models to handle long-range dependencies.

During the manual evaluation, we observed that the self-attention-based models produce fluent outputs for the sentence dataset. The fluency drops for the sequence to sequence model when trained on the paragraph dataset. However, the fluency score of the Transformer model increases slightly. The fidelity scores follow a similar pattern exhibited by the fluency scores. The Transformer model performs well even in the paragraph dataset, reconfirming its ability to model long-range dependencies.

Although the results are usable in many cases, we observe three types of deviation from the expectations for handwritten titles:

Lack of company specific wording: Some companies use their names in the headers (titles), for example, *Microsoft internet explorer*, *will BBC share your information?*, and *NSF subscription management*. These titles are hard to train, and they are harder to generate if they belong to the test set. The model outputs junk and non-fluent titles for these cases.

Brevity when the reference text is longer: All the models fail to generate relevant outputs when there are lengthy reference responses. Long titles such as *we will share information with our partners and affiliates*, *notes regarding the use of the website by children*, and *how does this apply to European Union and Swiss residents?* are difficult to reproduce, and the output quality decreases with an increase in its length.

Problems with extremely short and long inputs: Lengthy inputs, especially text greater than eight sentences, are challenging to train as they require high GPU memory and larger models. This partially explains the dip in the performance of the paragraph dataset when compared to the sentence dataset. On the other hand, concise sentences contain very little information to generate relevant topics resulting in unigram titles that defeat the purpose of a generative model.

6 Conclusion

One of the primary concerns of privacy policies is its length, which leads to readability issues. We created a Transformer-based automatic title generation system that makes clever use of the document structure of web privacy policies to generate short but meaningful titles for all paragraphs of a privacy policy. These titles act as micro summaries to various sections of the policy, opening the doors for better organization of privacy policies. Further research is required to evaluate the usefulness of titles, such as asking users to answer representative sets of privacy questions with and without the benefits of the approach presented in this paper.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. #CNS-1914444.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. https://arxiv.org/pdf/1409.0473.pdf.

[2] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018. https://arxiv.org/pdf/1803.11175.pdf.

[3] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. https://www.aclweb.org/anthology/D14-1179.pdf.

[4] Hyo Shin Choi, Won Sang Lee, and So Young Sohn. Analyzing research trends in personal information privacy using topic modeling. *Computers & Security*, 67:244–253, 2017. https://www.sciencedirect.com/science/article/pii/S0167404817300603.

[5] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. http://www.ric.edu/faculty/organic/coge/cohen1960.pdf.

[6] Federal Trade Commission et al. Protecting consumer privacy in an era of rapid change. *FTC report*, 2012. https://journalprivacyconfidentiality.org/index.php/jpc/article/download/596/579.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. https://arxiv.org/pdf/1810.04805.pdf.

[8] Desmond Elliott and Frank Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, 2014. https://www.aclweb.org/anthology/P14-2074.pdf.

[9] Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. Self-attentive model for headline generation. In *European Conference on Information Retrieval*, pages 87–93. Springer, 2019. https://arxiv.org/pdf/1901.07786.pdf.

[10] Sebastian Gehrmann, Steven Layne, and Franck Dernoncourt. Improving human text comprehension through semi-markov crf-based neural section title generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1677–1688, 2019. https://arxiv.org/pdf/1904.07142.pdf.

[11] Abhijith Athreya Mysore Gopinath, Shomir Wilson, and Norman Sadeh. Supervised and unsupervised methods for robust separation of section titles and prose text in web documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 850–855, 2018. https://www.aclweb.org/anthology/D18-1099.pdf.

[12] Yuko Hayashi and Hidekazu Yanagimoto. Headline generation with recurrent neural network. In *New Trends in E-service and Smart Computing*, pages 81–96. Springer, 2018. https://link.springer.com/chapter/10.1007/978-3-319-70636-8_6.

[13] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3(1):25, 2016. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0059-y.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. https://arxiv.org/pdf/1412.6980.pdf.

[15] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, jul 2004. Association for Computational Linguistics. https://www.aclweb.org/anthology/W04-1013.

[16] Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *2016 AAAI Fall Symposium Series*, 2016. https://www.aaai.org/ocs/index.php/FSS/FSS16/paper/download/14099/13701.

[17] Shuang Liu, Renjie Guo, Baiyang Zhao, Tao Chen, and Meishan Zhang. Appcorp: A corpus for android privacy policy document structure analysis. *arXiv preprint arXiv:2005.06945*, 2020. https://arxiv.org/pdf/2005.06945.pdf.

[18] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *ISJLP*, 4:543, 2008. https://kb.osu.edu/bitstream/handle/1811/72839/1/ISJLP_V4N3_543.pdf.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. http://www.cs.cmu.edu/~jeanoh/16-785/papers/papineni-acl2002-bleu.pdf.

[20] David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. Unsupervised topic extraction from privacy policies. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 563–568, 2019. https://dl.acm.org/doi/pdf/10.1145/3308560.3317585.

[21] Natalie Schluter. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, 2017. https://www.aclweb.org/anthology/E17-2007.pdf.

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[23] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. From neural sentence summarization to headline generation: A coarse-to-fine approach. In *IJCAI*, pages 4109–4115, 2017. https://www.ijcai.org/Proceedings/2017/0574.pdf.

[24] Ottokar Tilk and Tanel Alumäe. Low-resource neural headline generation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 20–26, 2017. https://www.aclweb.org/anthology/W17-4503.pdf.

[25] Ngoc Tran, Hieu Tran, Son Nguyen, Hoan Nguyen, and Tien Nguyen. Does bleu score work for code migration? In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*, pages 165–176. IEEE, 2019. https://dl.acm.org/ft_gateway.cfm?id=3339104&type=pdf.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[27] 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, Thomas B Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 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)*, pages 1330–1340, 2016. https://www.aclweb.org/anthology/P16-1126.pdf.

[28] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 617–626, 2018. https://dl.acm.org/doi/pdf/10.1145/3269206.3271711.