Towards Private Synthetic Text Generation

Rishi Bommasani
Department of Computer Science
Cornell University
rb724@cornell.edu

Zhiwei Steven Wu
Computer Science
Engineering Department
University of Minnesota
zsw@umn.edu

Alexandra Schofield
Department of Computer Science
Harvey Mudd College
xanda@cs.hmc.edu

Abstract

The integration of privacy guarantees in machine learning is crucial to modeling sensitive data. Recent work has approached this problem by adjusting machine learning algorithms over sensitive data in order to satisfy theoretical privacy guarantees. However, this approach restricts the iterative process of refining model choice and parameters. An attractive alternative is to generate synthetic data to serve as a drop-in, reusable replacement for the sensitive data. However, in the important setting of text, existing private generative models do not adequately guarantee the recreation of properties of natural language in the original corpus that might be of interest. In this work, we propose the fundamental task of private synthetic text generation. The goal is to construct a synthetic corpus that satisfies a provable privacy guarantee while both replicating distributional properties of the original data and resembling natural language. We argue these three conditions are pivotal for privacy-preserving natural language processing, and discuss potential challenges and evaluation frameworks that afford flexibility for tighter specifications in future work. Though this task is challenging, we offer some initial approaches that can explicitly leverage public data to improve the quality of the synthetic output.

1 Introduction

Large-scale analysis of human-generated text data can provide insights in applications ranging from search and recommendation to medicine, social sciences, and even literature. However, the diversity of expression available in text also furnishes a diversity in the types of sensitive information that text analyses may leak, such as personally identifying information or private narratives.

In recent years, *differential privacy* (Dwork et al., 2006) has emerged as a dominant theoretical framework to provide concrete privacy guarantees when performing computation over sensitive data collections. In response to interest in precisely these kinds of sensitive data distributions, researchers have integrated differential privacy into numerous machine learning frameworks (Abadi et al., 2016; Team et al., 2017), including in applications for textual data (Park et al., 2016; Zhu et al., 2016; McMahan et al., 2018). Within both privacy-preserving machine learning and privacy-preserving natural language processing, the standard approach for integrating privacy has been through *randomized response* (Warner, 1965), or random noise-based perturbations of intermediate computations within a given machine-learning model.

NeurIPS 2019 Workshop on Machine Learning with Guarantees, Vancouver, Canada.

This algorithm- or model-centric view of privacy has several drawbacks, especially with respect to text. First, present methods of introducing privacy are often bespoke to a particular machine learning algorithm, requiring complex proofs to layer into iterative and stochastic convergence algorithms. Second, if inference of a model that spent some initial privacy budget reveals bugs or undesirables in that model, a new version of that model further expends the privacy budget. Third, many of these methods are poorly suited for the sparsity and high-dimensional nature of text, such that a reasonable privacy budget will remove all but the most basic patterns of text frequency.

An alternative approach that largely addresses the first two issues is *private synthetic data generation*. In this approach, one uses a privacy-preserving generative model to construct synthetic data that simulates the original collection. Because this process spends a fixed privacy budget only once to generate the synthetic dataset, subsequent analyses/models of the released data do not spend more privacy budget, which permits arbitrary exploration and testing of models. While it may be impossible to entirely reconstruct the distributional properties of the original sensitive data, private synthetic data generation approaches allow the initial introduction of differential privacy into novel domains.

However, existing methods for private synthetic data generation are unlikely to work in the domain of text if directly applied to text with no further modifications due to the large number of sparse features. In response to this challenge, we introduce private synthetic text generation as an integral step towards unifying theoretically robust notions of privacy, and empirically meaningful approaches for natural language. We propose that successful private synthetic data generators (a) are provably private, (b) preserve distributional trends, and (c) generate grammatical text. Systems that jointly satisfy these three properties can generate data to enable subsequent NLP practitioners to analyze and model the sensitive data as they choose while preserving privacy. We consider the inherent challenges of this task, and give an evaluation schema that provides results that are well-correlated with potential future analyses/models. Finally, we contribute two initial methods towards this task that leverage public data to help improve generation quality in this challenging private setting.

2 Framework

The demand for private approaches towards natural language processing spans domains such as medical text and legal documents to text messages and private communications. Emergent technologies, such as voice assistants, introduce further sensitive data about users that must be protected when used in training of public-facing models. For this diverse class of text, we anticipate that synthetic corpora that incorporate statistical aspects of text collections of interest could replace the original sensitive text for data consumers performing data analyses or learning unsupervised models.

We require three criteria for successful private synthetic text generation:

1. **Privacy:** This is inherently necessary to provably and properly safeguard the sensitive data.
2. **Distributional Similarity:** The results of the analysis/learned model are only meaningful if they reflect behavior consistent with the original sensitive data.
3. **Text:** Existing analysis methods and unsupervised models assume the natural language input is presented as text. While some models (Blei et al., 2003) may only require count statistics, other models may rely on sequential and syntactic information for feature extraction.

Trading off these three demands is immensely challenging. Natural language generation remains difficult to implement (Rush, 2019), and many existing systems are exceptionally brittle, especially in the presence of noise (Belinkov & Bisk, 2018). Further, existing methods for natural language generally need to only optimize for either textual quality (Sutskever et al., 2014; Paulus et al., 2018) or distributional similarity (Blei et al., 2003) but not both. We anticipate that the approach to find the appropriate balance of these three aspects may be particular to the domain of the data and application: for instance, when clustering medical records, a model that preserves the frequencies of rare unigram features may help less than one that preserves frequent higher-order n-grams.

A technical problem we foresee in the private setting is that sensitive datasets are generally of a fixed size too small to feed the data-hungry approaches that are prevalent in natural language generation at present. To address this, we decompose the process of construction of the synthetic corpus into a generative step followed by a sampling procedure where at least one of the two pieces make use of public data. In §4, we discuss two initial models that fit this form. The approach we describe is

therefore not end-to-end, which risks poor performance from cascading errors. However, we argue that the two-stage nature serves an important purpose in allowing for the "sourcing" of how the public and private data ultimately contribute to the synthetic data we produce. As we discuss in §4, this may be of particular value given the deficiencies we see in our initial approaches.

3 Evaluation

Since our framework expects a provable privacy guarantee, no further evaluation of how well privacy is preserved is explicitly required beyond verification that the privacy guarantee was proven correctly. However, we advocate consideration of multiple privacy settings in evaluating the other two components to find one best suitable for the given data domain.

The natural language processing literature furnishes a number of options to evaluate distributional similarity of text collections. We suggest focusing on distributional properties that correlate well with the information we expect to be most useful in subsequent analyses/models. Concretely, if we anticipate that topic models will be applied to the synthetic data to analyze particular salient trends, prioritizing unigram-level distributional similarity may suffice as many topic models are insensitive to word-order. On the other hand, to learn contextual word embeddings, higher-order n-gram statistics may be more relevant to capture the local contexts of words correctly.

In considering potential evaluation strategies for the quality of the generated text, we note that pretrained language models (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019) can be a lightweight and straightforward estimator of textual quality. In particular, recent results (Chowdhury & Zamparelli, 2018; Warstadt et al., 2018) suggest that neural language models are surprisingly promising approximators of human judgments regarding grammaticality and acceptability.

4 Models

We detail two initial model designs to execute private synthetic text generation. We intentionally do not completely specify all aspects within our private synthetic text generation framework. Freedom to apply different generative approaches and privacy definitions in this framework is useful both to encourage creative approaches to the task and refinement within the framework for particular domains. Within our initial exploration of models that can perform this challenging task, we consider two approaches that both contain a generative step followed by a sampling step. In each case, since the generative model is learned privately, the subsequent sampling retains the privacy guarantee due to the *Post-Processing Theorem* (Dwork & Roth, 2014). For both approaches, one of these steps leverages public data via pretrained language models (Devlin et al., 2019) given their tremendous recent successes in modelling natural language.

Approach 1. In our first approach, we begin by learning a private topic model (Park et al., 2016) on the sensitive data that is inferred using variational inference. Given the learned private topic model which satisfies document-level (ε, δ)-differential privacy computed using moments accountant (Abadi et al., 2016), we can sample bag-of-words vectors by executing the generative LDA process: sample a categorical distribution θ over topics from the Dirichlet document-topic prior and a Poisson document length d, then sample d individual words independently from the induced distribution θ. This produces a synthetic collection of bag-of-words vectors that we find is reliably distributionally similar to the original corpus at the unigram level and that is guaranteed to be privacy-preserving.

We then find an ordering of the bag-of-words vector by searching the space of orderings or permutations, using beam-search to make this computationally tractable and using GPT-2 (Radford et al., 2019) to score partial word sequences (Schmaltz et al., 2016). We note that such an approach is a strict improvement over using the private topic model as a private synthetic text generator (as we get n-grams beyond unigrams). The resulting generation is distributionally similar for bigrams and trigrams and scores well under a pretrained language model. Qualitatively, we also observe that the generations seem to regularly abide by shallow syntactic constraints. However, we also see a dramatic performance decline (in both text quality and distributional similarity) for generating longer documents and for more stringent privacy settings. In inspecting the behavior, we source these issues to the private topic model: the inferred topics degrade rapidly for increased privacy, and sampled bag-of-words vectors cannot consistently be assembled into natural sentences for increased

sentence lengths. In general, while this method may be suitable for private synthetic text generation for short texts (such as text messages), we do not believe it will extend to longer documents due to the word-ordering task being inherently ill-posed for these settings.

Approach 2. In our second approach, we again consider how to use public data and pretrained language models but instead consider shifting the information into the initial generative phase of our two-step pipeline. Similar to the first method, we also consider how we can further improve an existing private generative model that is appropriate for text. Concretely, we begin with a pre-trained GPT-2 model (Radford et al., 2019), then fine-tune it on the sensitive data with a word-level privacy guarantee using a differentially private analogue of stochastic gradient descent (SGD) (Song et al., 2013; Abadi et al., 2016), where privacy loss is again accounted for using the moments accountant. This approach takes inspiration from *inoculation by fine-tuning* (Liu et al., 2019), though in this method, we demonstrate *privatization by fine-tuning*. Empirically, we find the behavior of this method is tightly tied to the learning rate used during fine-tuning. When we fine-tune conservatively (in accordance with standard NLP practices), we observe that we largely do not learn the sensitive data distribution and distributional similarity scores are low. Conversely, when fine-tuning is aggressively conducted, we do note slight increases in distributional similarity with respect to the source data, but we largely lose the pretraining benefits towards textual quality. More sophisticated learning rate schedules, especially adjusted to using a differentially private optimizer, may be required for this synthetic text generation approach to effectively leverage both pretraining and the sensitive data.

5 Future Directions

Given our initial findings, we identify several directions for future inquiry. Both of the private mechanisms we consider are entirely unaware of the injected noise (i.e. the algorithms were privatized by randomized response and inference behaves as it would in the non-private setting). As Schein et al. (2019) observes, we may see improved performance by adjusting the inference algorithms to account for this fact. Additionally, in both of our models, our statement of the privacy guarantee is with respect to (ε, δ)-differential privacy. However, under our framework, any theoretical notion of privacy (He et al., 2014; Mironov, 2017; Geumlek & Chaudhuri, 2019; Schein et al., 2019) can be used so long as it is robust to post-processing. Other privacy definitions may be more suitable for text and/or for particular domains of interest. Beyond this, we believe that a more seamless integration of public and private data is integral to better performance on the generation aspects of this task. While our approaches separate the two, which is beneficial for understanding which data and model component is contributing to poor performance in some settings, we expect that an underlying concern is domain adaptation (Zhou et al., 2016). Introducing public data risks domain shift between the pretraining corpora and the sensitive corpora (Ruder, 2019). Language model fine-tuning may be a simple solution for this as recent work (Han & Eisenstein, 2019) demonstrates. This further emphasizes the value of improved understanding regarding fine-tuning mechanisms, learning rate scheduling, and their interplay when employing differentially private SGD.

6 Conclusion

We propose the fundamental task of *private synthetic text generation*, which we believe to be central to the process of better integrating theoretical privacy and natural language processing. We provide a flexible but concrete framework for this task that enumerates three criteria that are integral to successful private synthetic text generators. We also discuss evaluation protocols and initial methods towards this task. Taken together, we hope this encourages further study of the challenging problem of private synthetic text generation to ensure provable privacy in NLP applications.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJ8vJebC-.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

Shammur Absar Chowdhury and Roberto Zamparelli. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 133–144, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, volume 3876, pp. 265–284, 2006.

Joseph Geumlek and Kamalika Chaudhuri. Profile-based privacy for locally private computations. *CoRR*, abs/1903.09084, 2019. URL http://arxiv.org/abs/1903.09084.

Xiaochuang Han and Jacob Eisenstein. Unsupervised domain adaptation of contextualized embeddings: A case study in early modern english. *CoRR*, abs/1904.02817, 2019. URL http://arxiv.org/abs/1904.02817.

Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility tradeoffs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pp. 1447–1458, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2376-5. doi: 10.1145/2588555.2588581. URL http://doi.acm.org/10.1145/2588555.2588581.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. 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)*, pp. 2171–2179, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1225. URL https://www.aclweb.org/anthology/N19-1225.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private language models. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Mijung Park, James R Foulds, Kamalika Chaudhuri, and Max Welling. Private topic modeling. In *Proceedings of the NeurIPS Workshop on Private Multi-Party Machine Learning*, 2016.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkACllQgA-.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

Alexander M. Rush. Pretraining for generation. June 2019. URL http://nlp.seas.harvard.edu/slides/Pre-training%20for%20Generation.pdf.

Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, and Hanna Wallach. Locally private bayesian inference for count models. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Allen Schmaltz, Alexander M. Rush, and Stuart Shieber. Word ordering without syntax. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2319–2324, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1255. URL https://www.aclweb.org/anthology/D16-1255.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 245–248. IEEE, 2013.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Apple's Differential Privacy Team et al. Learning with privacy at scale. *Machine Learning Journal*, 1, 2017. URL https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *CoRR*, abs/1805.12471, 2018. URL http://arxiv.org/abs/1805.12471.

Guangyou Zhou, Zhiwen Xie, Jimmy Xiangji Huang, and Tingting He. Bi-transferring deep neural networks for domain adaptation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 322–332, 2016.

Tianqing Zhu, Gang Li, Wanlei Zhou, Ping Xiong, and Cao Yuan. Privacy-preserving topic model for tagging recommender systems. *Knowledge and Information Systems*, 46(1):33–58, 2016.