(ε, δ)-Differentially Private Text Generation via Generalizing Named Entity to Reduce Utility Loss

Taisho Sasada

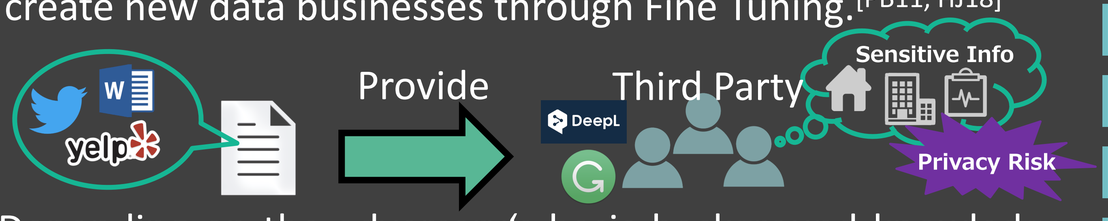
mail : sasada.taisho.su0@is.naist.jp

Abstract

- Since an anonymization assume the knowledge possessed by the adversary, sensitive information may be leaked depending on the adversary's knowledge even after this anonymization. Therefore, instead of providing original data, a method to generate differentially-private synthetic data has been proposed. Differential privacy (DP) is more flexible than anonymization technologies because it does not require the assumption of the adversary's knowledge. However, if a large noise is added to the gradient in text generative model to satisfy DP, the utility of the synthetic text is degraded. Since DP can be satisfied with a small noise in data containing duplicates, it is possible to reduce utility loss as text by creating duplicates before adding noise. In this study, we reduce the amount of noise added by creating duplicates through generalization, thereby minimizing text utility loss. By constructing a differentially-private text generation model, we can provide synthetic text and promote text utilization while protecting privacy information in the text.

1. Background & Motivation

- Increasing demand for providing collected user-generated text to third parties to build machine learning models and create new data businesses through Fine Tuning.[PB11, HJ18]



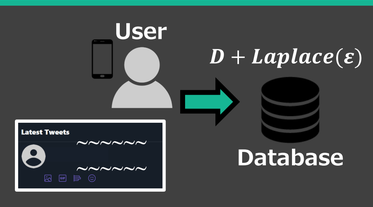
- Depending on the adversary‘s basic background knowledge, sensitive information may be inferred.

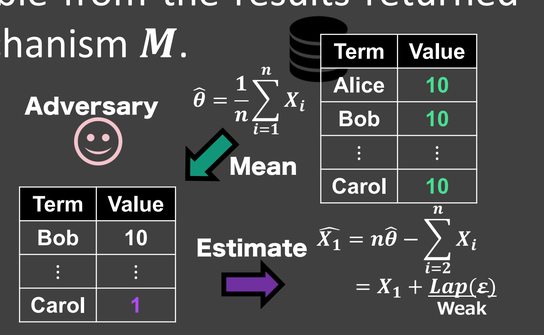
[PB11] Park, Byung-Kwon, and Il-Yeol Song. "Toward total business intelligence incorporating structured and unstructured data." In Proceedings of the 2nd International Workshop on Business intelligencE and the WEB, pp. 12-19. 2011.
[HJ18] Howard, Jeremy, and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 328-339. 2018.

2. Preliminary

- (ε, δ)-DP[DC06] is protect privacy in the output because there is less information available from the results returned through the randomization mechanism M.

$$\text{DP}: Pr[M(D) \in y] \leq \exp(\varepsilon) \cdot Pr[M(D') \in y]$$



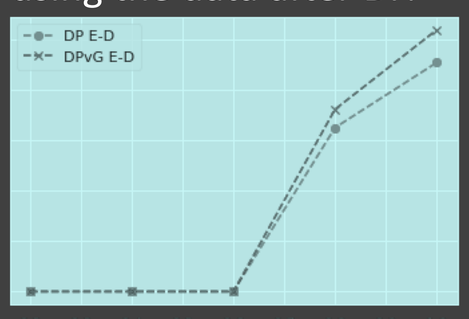


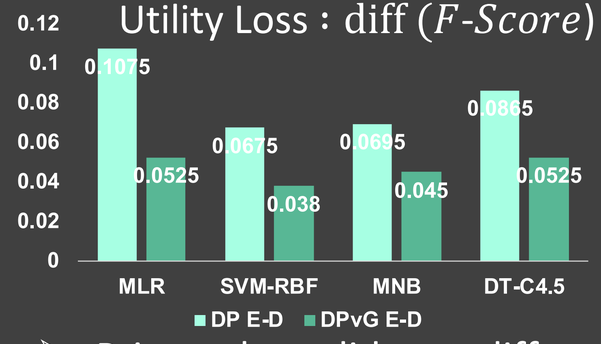
- By satisfying k-anonymity and then adding noise, DP can be satisfied with a weak noise[SJ14].

[DC06] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis." In Theory of cryptography conference, pp. 265-284. Springer, Berlin, Heidelberg, 2006.
[SJ14] Soria-Comas, Jordi, Josep Domingo-Ferrer, David Sánchez, and Sergio Martínez. "Enhancing data utility in differential privacy via microaggregation-based k-anonymity." *The VLDB Journal* 23, no. 5 (2014): 771-794.

3. Experimental Result

- We evaluated Utility Loss by the difference between the results using the data before DP and the results using the data after DP.





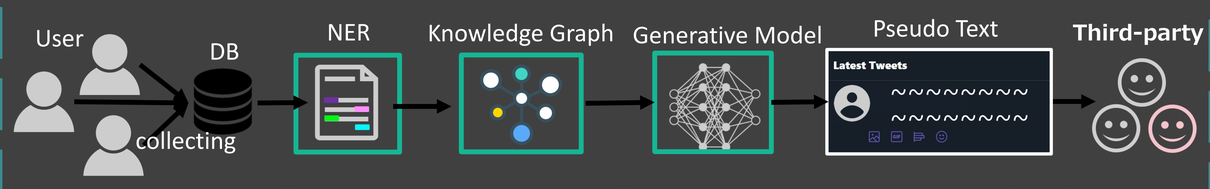
- Privacy loss did not differ significantly from the base model, which indicate that data privacy is sufficiently protected.

Balancing privacy protection and limiting loss of usefulness

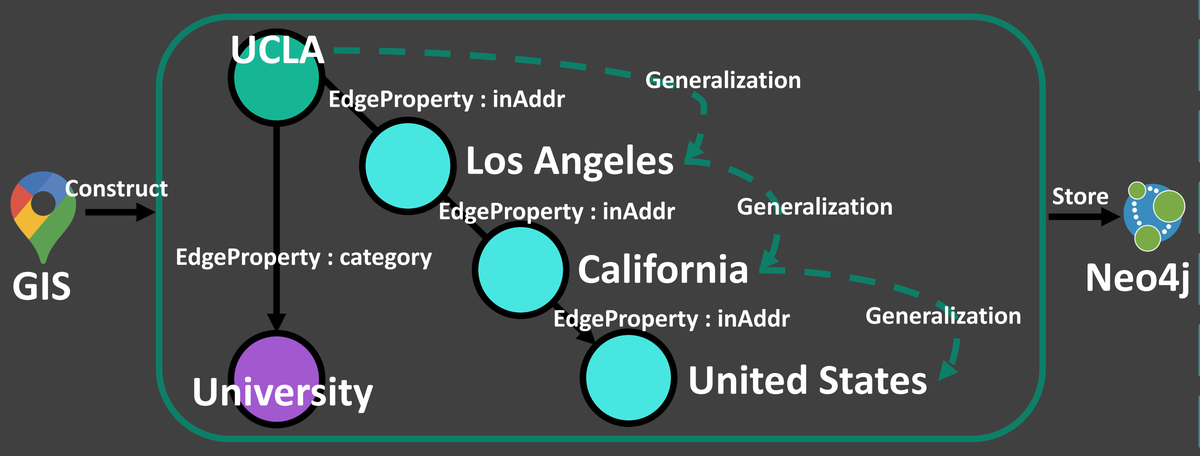
[MA11]Maas, Andrew, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. "Learning word vectors for sentiment analysis." In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142-150. 2011.

2. Proposed Method

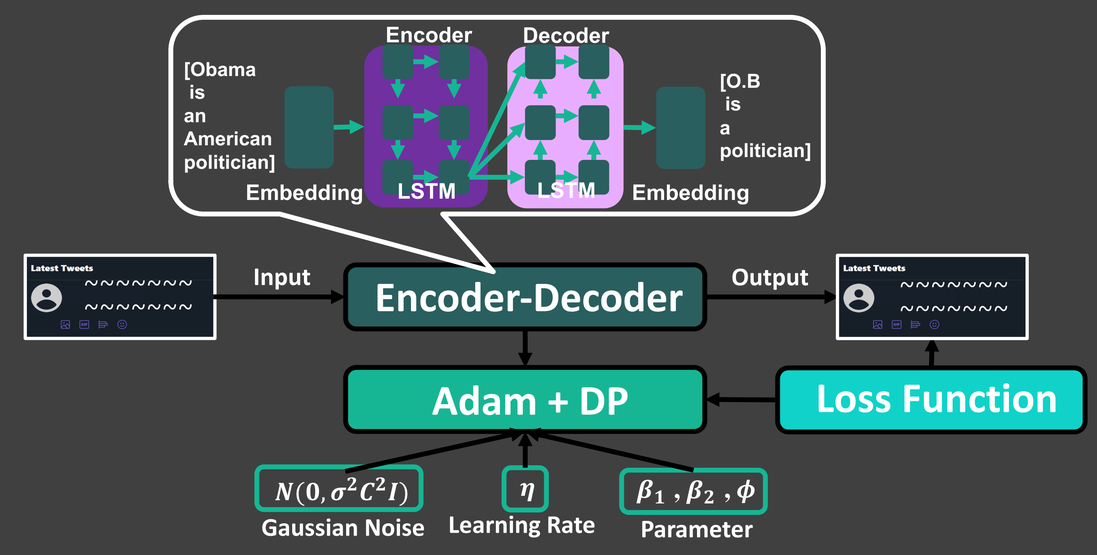
- BERT is used to extract named entities. Since the named entities retain the conceptual form, we can correctly substitute them for higher-level words by generalization.



- We generalize unique named entities to create duplicates satisfying k-anonymity, and reduce the amount of noise added, and suppress utility loss.



- By mediating the generative model, it is possible to provide privacy-preserved pseudo-text.



[SJ16] S. Mishra and J. Diesner. Semi-supervised Named Entity Recognition in noisy-text. In Proceedings of the 2nd Workshop on Noisy User-generated Text, pages 203–212, 2016.
[S02]Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): 557-570.

4. Conclusion and Future Work

- We propose a framework for providing useful text while protecting privacy by constructing a DP generative model.
- Compared with base model, our framework improve data utility loss.
- In future work, our generative model must require to extension toward conditional generative model.