E-045

単語レベル差分プライバシーの日本語テキストへの適用とその特性の検証
Application of Word-Level Differential Privacy to Japanese Text and Verification of Its Properties

前田 佑斗† 安藤 一秋‡
Yuto Maeta Kazuaki Ando

1. はじめに

大量のデータが生成され，蓄積される昨今，新事業やサービスの改善のためにデータを外部機関に提供して分析に役立てるなどのニーズが高まっている．2015 年の個人情報保護法の改正により，特定の個人を識別できないようにする匿名加工を適用すれば，本人の同意を得ていない場合でも，個人情報を含むデータを第三者に提供することが可能になった．構造化データに対しては，十分に利用可能な匿名化手法が提案されているが，テキストデータのような非構造化データに対しては確立された手法が現状存在しない．

本稿では，5 種類の単語レベル差分プライバシーメカニズムを日本語テキストに適用し，それらの特性を検証する．

2. 関連研究

日本語テキストに対する匿名化として，荒巻らは，最長共通部分列の長さに着目し，テキストに対する k-匿名化手法[1]を提案した．また，前田らは，文字 n-gram を用いて，共通部分文字列を置換する k-匿名化手法[2]を提案した．さらに，清水らは，エンティティの k-匿名化後に，大規模言語モデルで匿名化した文を生成する手法[3]を提案した．清水らの手法では，エンティティ系列への変換によって構造化データに変換することを可能にしている．

英語テキストに対する匿名化として，Xu らが英語からフランス語，さらに英語に翻訳する逆翻訳を意識した匿名化手法[4]を提案した．また，Meisenbacher らは，単語ベースの差分プライバシーの各種手法について，包括的に比較分析[5]した．しかし，日本語においてどの程度の有効性を持つのか，また，どのような特性を示すのかについては未だ十分に検証されていない．

差分プライバシーを自然言語処理に応用する研究は，単語埋め込みベクトルの摂動以外にも，モデル訓練時の勾配にノイズを加える DP-SGD[6]といったアプローチも存在し，広範な研究領域を形成している．本稿では，これらの多様なアプローチの中でも，特に Meisenbacher らが英語テキストで検証した単語レベルの差分プライバシー手法[5]に着目し，日本語テキストへの適用可能性と，その際のプライバシー・ユーティリティのトレードオフ特性を明らかにする．

3. 実験設定

3.1 データセット

本稿では，データセットとして，TIS 株式会社の chABSA-dataset[7]を用いる．chABSA-dataset は，上場企業の有価証券報告書の各文に対してネガティブ・ポジティブの感情を付与した感情分類用データセットである．chABSA-dataset は文レベルではなく，単語レベルでアノテーションされている．そこで，文内の単語に対するアノテーション数をもとに，任意の 1,000 文に対してアノテーション処理を施したものをデータとして利用する．また，テキストからテキストに変換する差分プライバシー技術では，近似する単語とそのベクトルが必要となる．そのため，日本語 Wikipedia データからランダムにテキストを抽出し，形態素解析器（Mecab+ipadic）で単語分割し，日本語 text8 ファイルを作成する．このファイルから語彙を抽出し，GloVe，Word2Vec，fastText を用いて，50 次元，100 次元，300 次元の単語分散表現を獲得して利用する．

3.2 採用する差分プライバシーメカニズム

Meisenbacher らの論文[5]を参考に，本稿では，以下の 5 種類の差分プライバシーメカニズムを利用して，日本語テキストにおける特性を明らかにする．

- Calibrated Multivariate Perturbations (CMP)
- Mahalanobis Mechanism
- Truncated Gumbel Mechanism
- Vickrey Mechanism
- Truncated Exponential Mechanism (TEM)

3.3 実験方法

本稿では，日本語テキストにおける差分プライバシーの有効性を多角的に評価するため，ユーティリティとプライバシーの指標を導入する．ユーティリティ指標には，正解率を利用する．なお，文献[5]を参考に LightGBM を用いた感情分析において 10 分割交差検証で算出，評価する．プライバシー指標は，単語の再識別リスクを測る推測成功率で評価する．

4. 実験結果

実験結果を図 1 と図 2 に示す．図 1 は，各分散表現，各次元でプロットしており，縦軸は正解率，横軸はプライバシー強度 ε の値（値が小さいほどノイズが大きいことを示す値）であり，右に行くほど ε の値が大きく，つまり，プライバシーの効用が低くなっている．図 2 は，50 次元の GloVe 埋め込みを用いた場合の結果であり，縦軸を正解率，横軸を推測成功率としたプライバシー・ユーティリティのトレードオフを示す．

図 1 より，Truncated Exponential 手法（TEM）の Accuracy が最も高くなり，匿名化を施す前のテキストに対する結果（ベースライン）との性能差が小さいことを確認した．さらに，プライバシー指標である推測成功率を評価した結果，図 2 に示すように，TEM は他の手法に比べて高い値を示し，ユーティリティを維持する代償として再識別リスクが高まる可能性が示唆された．

† 香川大学大学院創発科学研究科　Graduate School of Science for Creative Emergence, Kagawa University
‡ 香川大学創造工学部　Faculty of Engineering and Design, Kagawa University

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

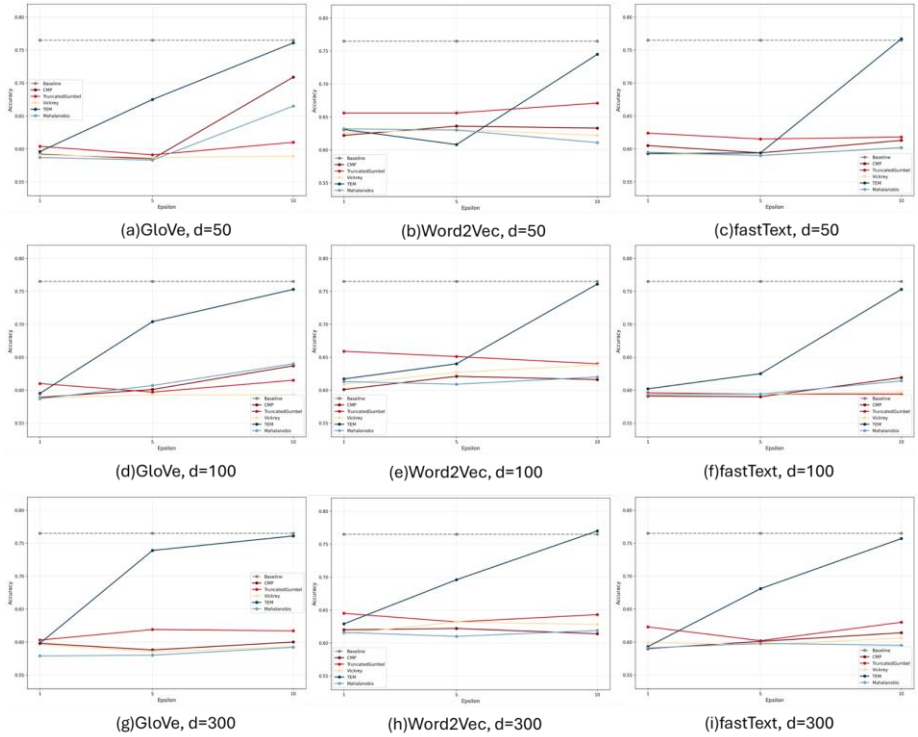

図 1　感情分類タスクにおける正解率と単語分散表現，分散表現の次元

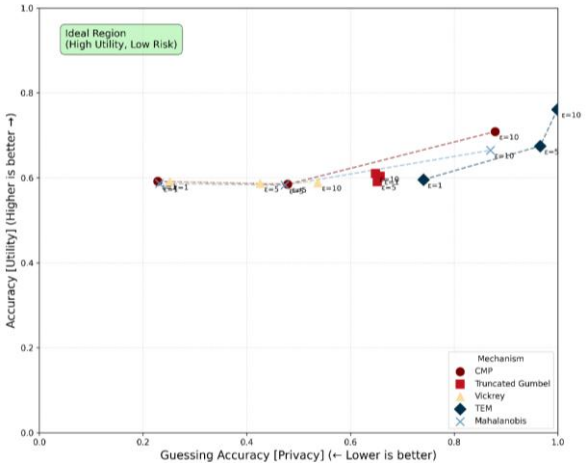

図 2　Glove，50 次元における推論成功率とユーティリティ

図 1 より，Vickrey 手法は，いずれの分散表現および次元においても，ε が 10 と 1 の場合に最も Accuracy 差が小さいことを確認した．また，単語の埋め込み手法を変更した場合，分類性能に影響を及ぼすことがわかった．これは，単語分散表現の設計の違いが，単語ベクトルの探索時に影響を及ぼすことが要因だと考えられる．さらに，ε を固定して，分散表現の次元を変化させた場合は，Mahalanobis，CMP のいずれの手法においても正解率が低下することがわかった．これは，設計によるものであり，次元が増えるにつれてユーティリティが低下する可能性が推測される．

図 2 において，理想的な手法は，左上（高精度・低リスク）に位置するが，多くの手法は右上の領域に分布しており，日本語テキストにおいても両者の間に明確なトレードオフが存在することを確認した．今後は，検証を継続して，日本語テキストにおける有効性を明らかにするとともに，プライバシーとユーティリティを保つ手法を検討する．

5. おわりに

本稿では，差分プライバシーを用いた匿名化手法の日本語テキストに対する有効性を検証した．実験の結果，日本語テキストにおいても，プライバシーとユーティリティのトレードオフが著しいことを確認した．単語ベースの差分プライバシーを純粋に適用した場合，プライバシーとユーティリティのトレードオフが著しいため，根本的なアプローチを検討する必要がある．Mattern ら[8]が指摘するように単語レベルの差分プライバシーは，その特性上，利用できる分野の制約が大きい．差分プライバシーの定義上，アプローチの種類が多く，また，非構造化データへの応用に関しては検討の余地が大きいといえる．

今後は，前処理を含めて，より幅広い手法について検討する必要がある．また，自然言語処理における他のタスクについても適用することで有効性を検証する．

参考文献

[1] 荒巻他，“テキストの k-匿名化”, IPSJ DBS-155, pp.1-8, 2012.
[2] 前田他，“ソーシャルメディアにおける非構造化テキストデータの k-匿名化によるプライバシー保護”, IPSJ DBS-162, pp.1-8, 2015.
[3] 清水他，“RECORD TWIN:病歴を保ちつつ表現が異なる症例を生成する”, IPSJ NL-260, pp.1-10, 2024.
[4] Q. Xu, et al., “Privacy-Aware Text Rewriting”, Proc. of INLG2019, pp.247-257, 2019.
[5] S. Meisenbacher, et al., “A Comparative Analysis of Word-Level Metric Differential Privacy: Benchmarking the Privacy-Utility Trade-off”, Proc. of LREC-COLING 2024, pp.174-185, 2024.
[6] M. Abadi, et al., “Deep Learning with Differential Privacy”, Proc. of the 2016 ACM SIGSAC, pp.308-318, 2016.
[7] https://github.com/chakki-works/chABSA-dataset
[8] J. Mattern, et al., “The Limits of Word Level Differential Privacy”, Findings of NAACL 2022, pp.867-881, 2022.

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.