固有表現抽出と差分プライバシーを用いた二段階匿名化手法と裁判記録における評価

前田佑斗[1] 安藤一秋[2]
[1]香川大学大学院 創発科学研究科 [2]香川大学 創造工学部
{s25g364, ando. kazuazki}@kagawa-u. ac. jp

概要

近年，法務や医療ドメインにおいて，テキストデータの二次利用への期待が高まっている．しかし，裁判記録などの公的文書には，多角的かつ潜在的な個人情報が含まれており，安全な匿名化が課題となっている．従来から用いられてきた固有表現抽出（NER）に基づく匿名化手法は，氏名や住所といった既知の識別子の除去には有効であるが，文脈や文体の組み合わせで生じる再識別リスクを完全には排除できない．そこで本稿では，識別子を構造的に除去する NER と，数理的にプライバシー保証を与える差分プライバシー（DP）を組み合わせた二段階匿名化手法を提案する．評価実験では，TAB Corpus を用い，4 種の DP メカニズムを比較する．特に，プライバシー予算（ε）に対する各手法の感度や，言語品質への影響，および情報理論的指標に基づく「情報の絞り込み」の度合いに注目して考察する．

1 はじめに

近年，デジタルデータの大規模な蓄積に伴い，法務，医療，行政などの各分野において，テキストデータを活用した高度な分析や AI システムの構築へのニーズが急速に高まっている. 特に, 裁判記録は, 法学研究の深化やリーガルテックの発展において極めて価値の高い情報資源である．しかし，これらの文書には，氏名や住所といった直接識別子に加え，職業，身体的特徴，特定のライフイベントといった準識別子が多く含まれており，それらの適切な保護が強く求められている．

テキストの匿名化においては，これまで固有表現抽出（Named Entity Recognition: NER）を用いて人名や地名を伏せ字にする匿名化手法（De-identification）が主流であった．しかし，近年の研究では，表層的な記号を置換するだけでは，文脈情報や特有の言い回しを手掛かりとして元の個人を推論できる可能性が指摘されている[1]. また, 大規模言語モデル（LLM）の発展により，わずかな文脈情報から推論が可能になったことから，再識別リスクが増大している．

こうした課題に対し，数学的なプライバシー保証を提供する差分プライバシー（Differential Privacy: DP）をテキストに適用する試みが進められている[2]. DP は，適切に設計されたノイズを付加することで統計的な安全性を保障する枠組みである一方，単語レベルでの適用では文法構造を損ないやすく，データの有用性が大きく低下するというトレードオフ（ゼロサムの関係）が存在する[4]．特に，裁判記録のような文脈的連続性やナラティブが重要なデータにおいては，有用性を保ちつつ安全性を担保する高度な設計が不可欠である．

本研究では，実務的なプライバシー保護への応用を見据え，二段階の匿名化パイプライン手法について提案する．第 1 段階では，高精度な NER により直接識別子を構造的に除去し，第 2 段階では，残存するテキストに対して DP に基づく確率的な攪乱を施す．本稿では，提案手法がプライバシー予算（ε）の変動に対してどのような挙動を示すかを実証的に比較・検証し，複雑なナラティブをもつテキストにおける匿名化の実現可能性を検討する．

2 関連研究

2.1 テキスト匿名化の評価基盤

テキストデータの匿名化に関する従来の研究では，人名や住所などの固有表現を伏せ字にする De-identification の性能（Recall/Precision）が主眼に置かれてきた．しかし，Pilán らは，裁判記録のような複雑な文脈情報を持つテキストデータにおいて，単純な識別子の除去だけでは不十分であることを指摘し，実体レベルでの保護を評価対象とする Text

Anonymization Benchmark（TAB）を提案した[1]．本研究では，この評価基盤を採用することで，文脈依存の再識別リスクを考慮した匿名化性能を検証する．

2.2 単語レベル差分プライバシーのメカニズム

数学的なプライバシー保証をテキストデータに導入する手法として，単語埋め込みベクトルにノイズを付加する Metric Differential Privacy（Metric DP）が提案されている[2]．主要なメカニズムとして，多変量ラプラスノイズを用いる CMP[5], 再構成リスクを低減するために第 2 近傍との距離を考慮する Vickrey[6], 埋め込み空間の密度に応じてサンプリングを調整する Gumbel[7]，および TEM[8]などが挙げられる．Stephen らによるベンチマーク調査[3]では，Gumbel メカニズムが, プライバシー保護とテキスト有用性のバランスにおいて，高い性能を示したと報告されている．

2.3 既存手法の限界と本研究の独自性

Mattern ら[4]は, 単語レベルの DP を一律に適用する手法には，文章長が固定されるといった理論的制約に加え，文法構造の崩壊や適応的攻撃（Adaptive Attack）に対する脆弱性が存在することを実証した．

本研究は，先行研究で指摘された「単体手法の限界」を踏まえ，「NER による構造的匿名化」と「DP による確率的匿名化」を用いた，二段階匿名化手法を提案する．これにより，Mattern らが指摘した品質劣化を抑えつつ，TAB が定義する高度な匿名化（Anonymization）をどの程度実現できるか検証する．

3 匿名化手法

テキストデータに含まれる識別子を構造的に除去する手法と，数理的なプライバシー保証に基づき確率的に攪乱する手法を組み合わせた二段階の匿名化パイプライン手法を提案する．

3.1 第 1 段階：NER による構造的匿名化

第 1 段階では，テキスト内の直接識別子および顕著な準識別子を構造的に特定し，除去することを目的とする．本稿では，基盤モデルとして Longformer を採用する．Longformer は，局所的なアテンションとグローバルなアテンションを組み合わせることで, 裁判記録のような長大な文書に対しても，広範な文脈情報を考慮した固有表現抽出を可能とする．具体的には，人名（PERSON），場所（LOC），日付（DATETIME）などの特定のエンティティスパンを検出し，それらを抽象的カテゴリ名へ置換する．これにより，表層的な個人情報の漏洩を防止する．

3.2 第 2 段階：DP による確率的匿名化

第 1 段階の処理を経た後であっても，文体や残存する単語の組み合わせに基づき，個人が推論されるリスクは残る．そこで，第 2 段階では，単語の埋め込みベクトル空間に Metric DP を適用し，数理的なプライバシー保証に基づく確率的攪乱を施す．

本研究では，以下の 4 種類のメカニズムを対象とし，それぞれの特性を比較・検証する．

1. CMP(Calibrated Multivariate Perturbations)
   多変量ラプラス分布から抽出したノイズを埋め込みベクトルに加算する標準的な手法である．
2. Vickrey Mechanism
   最近傍単語と次近傍単語との距離関係を考慮し，単語の再構成リスクを抑制する手法である．
3. Gumbel Mechanism
   埋め込み空間における単語分布の密度に基づき, 指数メカニズムを用いて最適な代替単語を確率的にサンプリングする手法である．
4. TEM (Truncated Exponential Mechanism)
   探索空間を一定の閾値（γ）で制限（Truncation）した上で, 局所的な埋め込み密度に応じてノイズ量を調整し，有用性の維持を図る手法である．

本パイプライン処理では, 第 1 段階で識別子を「構造的に除去」し，第 2 段階で「数理的な保証」を付与することで，単語レベルで一律に DP を適用した際に生じる文法構造の崩壊を抑制し，文脈の維持とプライバシー保護の両立を図る．

4 評価実験

4.1 実験設定

評価データセットには，裁判記録を対象とした Text Anonymization Benchmark (TAB) Corpus [1]のテストセット 127 件を用いる．

本実験では, 以下の 3 つのアプローチを比較する．第一に，NER only は，TAB Corpus で調整された Longformer ベースの NER モデルを用いて，固有表現を[MASK]トークンへ置換する手法である（ER_{di} =1.000，ER_{qi} =0.916，WP_{di+qi} =0.850）．第二に，DP only は, 原文に対して直接, DP メカニズムを適用す

る手法である．第三に，NER+DP は，本研究の提案手法であり，NER により固有表現を[MASK]化した後，残存テキストに対して DP を適用する．

DP only および NER+DP については, CMP, Vickrey, Gumbel，TEM の 4 種類のメカニズムを評価対象とし，プライバシー予算 $\varepsilon \in \{1, 5, 10\}$ の 3 水準を設定する．よって本稿では，合計 4 メカニズム × 3ε ×2 アプローチ =24 条件で評価する（詳細は付録表 2 参照）．各メカニズムのパラメータは先行研究に従い, 埋め込みモデルには GloVe[i] 50 次元を用いる．

4.2 評価指標

本研究では，プライバシー保護性能，有用性の維持，および言語品質の 3 つの観点から評価する．

プライバシー保護性能の評価には, TAB Corpus で定義された Entity-level 指標[1]を用いる．ER_{di}（Entity Recall - Direct Identifiers）は，直接識別子（氏名，事件番号など）が完全に保護された実体の割合を示す．ER_{qi}（Entity Recall - Quasi Identifiers）は，準識別子（職業，日付，場所など）が保護された実体の割合を示す. 従来の token-level に基づく評価とは異なり，entity-level 評価では，同一実体に関連付けられたするすべてのメンションが完全に保護されているかどうかを判定する点に特徴がある．

有用性の評価には，TAB Corpus で定義された WP_{di+qi}（Weighted Precision）[1]を用いる．これは，言語モデル（GPT-2）で推定した各トークンの出現確率に基づき，自己情報量で重み付けした適合率である．この指標により，一般的な頻出語より，希少でかつ情報量の大きい語の保存を重視した有用性評価が可能となる．低い WP は多くの情報が削除されたことを，一方で高い WP は原文に近いことを示す．

言語品質の評価には，Perplexity（PPL）および相互情報量（MI）を用いる．PPL は，事前学習済み言語モデル GPT-2[ii] を用いて算出し, 匿名化後のテキストの流暢さを測定する．値が低いほど，自然な文章であることを示す．MI は，原文 X と匿名化後テキスト Z の間の相互情報量を zlib 圧縮に基づいて近似的に算出する. MI が高いほど原文の情報が多く保持されていることを示す一方，再識別リスクの潜在的な指標ともなり得る．

[i] https://nlp.stanford.edu/projects/glove/
[ii] https://huggingface.co/openai-community/gpt2

5 結果と考察

TAB Corpus を用いた二段階匿名化手法に関する評価実験の結果を示し，プライバシー保護性能，有用性および言語品質の観点から考察する．

5.1 匿名化性能と有用性の関係

表 1 に，NER only，DP only，および提案手法 NER+DP の性能を示す．なお，DP only および NER+DP の各値は，4 メカニズムの平均値である．

表 1　三段階の匿名化手法の性能

手法	ER_{di}	ER_{qi}	WP_{di+qi}	PPL
NER only	1.000	0.916	0.850	11.52
DP only	1.000	0.957	0.236	695.9
NER+DP	1.000	0.972	0.177	331.8

実験の結果，すべての手法および条件において，直接的識別子に対する保護指標 ER_{di} は 1.000 を達成した．これは，NER による構造的除去と DP による確率的攪乱により，人名や住所といった主要な直接識別子が原文から完全に匿名化されたことを示す．

準識別子の保護指標 ER_{qi} においては，手法間で明確な差が確認された．NER only では ER_{qi}=0.916 であったのに対し，DP only では 0.957，NER+DP では 0.972 を達成した．NER+DP は，NER only と比較して 0.056 ポイントの向上を示し，相対的には 6.1%の改善に相当する．これは第 1 段階の NER が準識別子を[MASK]へ置換し, 第 2 段階の DP が残存する文脈情報に確率的攪乱を与えることで，保護レベルが段階的に向上したためと考えられる．

言語品質の観点からは，手法間で顕著な差異が観察された．NER only は PPL=11.5 と極めて高い流暢性を示している．これは, 固有表現を[MASK]に構造的に置換する処理が文法構造を保持していることを示している．一方，DP only では PPL=695.9 と，NER only の約 60 倍に増加しており，確率的な単語置換が文法構造の整合性を著しく破壊していることが確認された．これに対して，提案手法 NER+DP は PPL=331.8 を達成し, DP only と比較して約半分の値に抑えられている．この結果は，提案手法が「保護を強めると文章が崩壊する」という従来のトレードオフを効果的に緩和していることを示す．図 1 に，PPL-ER_{qi} 空間における各手法の分布を示す．NER+DP（塗りつぶしマーカー）は，高い準識別子

保護性能と高い流暢性を意味する理想領域（図の左上）に位置しており，提案手法の特性が視覚的に確認できる．

5.2 DP メカニズムの感度分析

4 種類の DP メカニズムについて，言語品質（PPL）と情報保持（MI）の観点から特性を分析する（詳細は付録表 2 参照）．

Gumbel は，プライバシー予算 ε の変化に対して最も安定的な挙動を示した．特に，NER+DP では PPL が 156〜158 の範囲に収まり，ε の値によらずほぼ一定となった．この特性は，実運用上のパラメータチューニングの負担が小さい点で優れており，Meisenbacher ら[3]のベンチマーク研究とも整合している．

TEM は，高 ε 条件において極めて高い流暢性を示した．特に，NER+DP の$\varepsilon = 10$で，PPL=12.6 を達成した．これは閾値 γ により，原文に近い代替語が優先的に選択されるためであるが，一方で MI も高く（1483.3 bytes），情報保持と再識別リスクのトレードオフが顕著である．興味深い点として，TEM（DP only，$\varepsilon = 10$）の ER_{qi} は 0.941 と最も低い値を示したが，NER+DP では 0.961 まで回復しており，構造的な固有表現除去が TEM の弱点を補完していることが確認された．

CMP と Vickrey は，全体として中間的な挙動を示したが，特に $\varepsilon = 5$ において PPL が急激に悪化する非線形な特性が観察された（CMP：1608.7，Vickrey：1633.6）．これは，多変量ノイズの影響により，低頻度語への置換確率が増大したことが影響したと考えられる．よって，実運用では $\varepsilon = 5$ 付近の設定を避ける必要がある．

5.3 二段階匿名化の有効性

実験結果は，二段階匿名化における役割分担の有効性を強く支持している．DP only は，準識別子まで含めた高水準のプライバシー保護を目指した場合，強力なノイズ付与が必要となり，その結果として PPL が 900 を超えるまで悪化する傾向にあった．例えば，CMP（$\varepsilon = 1$，DP only）は ER_{qi}=0.961 と高い保護性能を示しているが，PPL=899.6 と極めて低い流暢性を示しており，実用的なデータ利活用の加点からは大きな制約となる．

一方，提案手法では，NER が「準識別子の構造的保護」を担い，DP が「文体や未知の属性の確率的攪乱」に専念するという役割を分担することで，DP の担う負担を大幅に軽減できる．その結果，$\varepsilon = 10$ のような低ノイズ設定においても高い保護レベルを維持でき，DP only と比較して約 2 倍の流暢性（PPL の約半減）を実現した．具体的には，CMP（$\varepsilon = 10$）で約 46%，Vickrey（$\varepsilon = 10$）で約 51%，Gumbel（$\varepsilon = 10$）で約 47%改善しており，複数の DP メカニズムに対して効果が確認された．

これらの結果は，構造的除去と数理的攪乱を組み合わせたハイブリッドアプローチが，実用的な匿名化において有効であることを示している．特に，法務・医療ドメインのような「高い保護レベル」と「実用的な品質」の両立が求められる領域において，提案手法の有効性は顕著である．

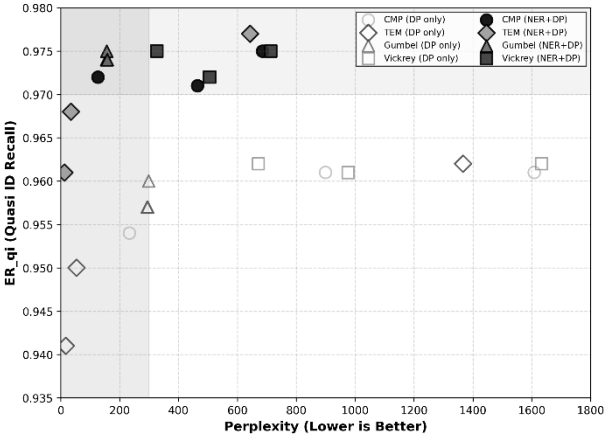

図 1　プライバシー保護性能と言語品質のトレードオフ

6 おわりに

本稿では，NER と DP を組み合わせた二段階匿名化手法を提案し，TAB Corpus を用いた実証評価により，その有効性を示した．実験の結果，NER と DP を組み合わせることで，単体手法の限界を相互補完できることを実証した．特に，DP のみを適用した場合に生じやすい文法構造の崩壊に対し，NER による事前処理が有効な緩和策になることを示した．また，4 種の DP の比較により，Gumbel の安定性，TEM の高感度性，CMP/Vickrey の非線形挙動といった特性を明らかにした．

本研究は，単語レベルの DP に焦点を当てており，文脈全体の意味情報を制御する点では限界がある．今後の課題としては，文レベル・段落レベルへの拡張，ドメイン適応型 DP メカニズムの設計が挙げられる．

参考文献

[1] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet, “The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization”, Computational Linguistics, pp. 1054–1101, 2022.

[2] Konstantinos Chatzikokolakis, Miguel E An drés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi, “Broadening the scope of differential privacy using metrics”, In Privacy Enhancing Technologies: 13th International Sym posium, PETS 2013, Proceedings 13, Springer, pp. 82–102, 2013.

[3] Stephen Meisenbacher, Nihildev Nandakumar, Alexandra Klymenko, Florian Matthes, “AComparative Analysis of Word-Level Metric Differential Privacy: Benchmarking The Privacy-Utility Trade-of”, LREC-COLING 2024, pp.178-185, 2024.

[4] Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum, “The Limits of Word Level Differential Privacy”, Findings of the Association for Computational Linguistics: NAACL 2022, pp.867-881, 2022.

[5] Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, Tom Diethe, “Privacy- and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations”, Proceedings of the 13th International Conference on Web Search and Data Mining, pp.178–186, 2020.

[6] ZekunXu, AbhinavAggarwal, OluwaseyiFeyisetan, Nathanael Teissier, “On a Utilitarian Approach to Privacy Preserving Text Generation”, Proceedings of the Third Workshop on Privacy in Natural Language Processing, pp.11-20, 2021.

[7] Nan Xu, private release of text Feyisetan, Abhi nav Aggarwal, Zekun Xu, Nathanael Teissier, “Density-Aware Differentially Private Textual Perturbations Using Truncated Gumbel Noise”, The International FLAIRS Conference Proceedings, 34, 2021.

[8] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, Ke Wang, “TEM: High Utility Metric Differential Privacy on Text” Proceedings of the 2023 SIAM International Con ference on Data Mining (SDM), pp.883-890, 2023.

A 評価データの詳細情報

評価に用いた詳細なデータおよび結果を表 2 に示す．表 2 では，4 メカニズム×3ε×2 アプローチ（DP only / NER+DP）の全 24 条件を示す.

表 2　適三段階の匿名化手法の性能比較

Mechanism	Epsilon	Type	ER_{di}	ER_{qi}	WP	PPL	MI
CMP	1	DP only	1	0.961	0.173	899.6	593.9
CMP	1	NER+DP	1	0.971	0.164	464.6	387.7
CMP	5	DP only	1	0.961	0.176	1608.7	668.3
CMP	5	NER+DP	1	0.975	0.164	685.4	453.6
CMP	10	DP only	1	0.954	0.219	233.5	1172.6
CMP	10	NER+DP	1	0.972	0.178	126.4	879.8
Gumbel	1	DP only	1	0.960	0.189	298.9	823.6
Gumbel	1	NER+DP	1	0.975	0.168	156.2	587.1
Gumbel	5	DP only	1	0.957	0.190	294	825.1
Gumbel	5	NER+DP	1	0.974	0.169	157.7	584.6
Gumbel	10	DP only	1	0.957	0.191	295.7	831.7
Gumbel	10	NER+DP	1	0.974	0.169	155.6	587.3
TEM	1	DP only	1	0.962	0.172	1367	606.8
TEM	1	NER+DP	1	0.977	0.161	642.7	406.5
TEM	5	DP only	1	0.950	0.313	53.9	1627.4
TEM	5	NER+DP	1	0.968	0.199	34.1	1246.7
TEM	10	DP only	1	0.941	0.674	17.5	1946.3
TEM	10	NER+DP	1	0.961	0.258	12.6	1483.3
Vickrey	1	DP only	1	0.961	0.173	976.6	593.3
Vickrey	1	NER+DP	1	0.972	0.164	505.6	388.2
Vickrey	5	DP only	1	0.962	0.175	1633.6	650.8
Vickrey	5	NER+DP	1	0.975	0.163	713.8	441.8
Vickrey	10	DP only	1	0.962	0.182	671.6	748.3
Vickrey	10	NER+DP	1	0.975	0.164	326.7	518.5

注：ER_{di}（直接識別子保護率），ER_{qi}（準識別子保護率），WP（加重適合率）， PPL（Perplexity），MI（相互情報量，単位：bytes）