DEIM Forum 2021 G31-4

プライバシ保護と内在情報価値保存を両立する擬似テキスト生成モデルの検討

笹田　大翔† 河合　将隆† 妙中　雄三† Doudou Fall† 門林　雄基†

† 奈良先端科学技術大学院大学先端科学技術研究科　〒 630-0192 奈良県生駒市高山町 8916-5

E-mail: †{sasada.taisho.su0,kawai.masataka.kl6,yuzo,doudou-f,youki-k}@is.naist.jp

あらまし　新規の研究やマーケティング等におけるデータ利活用の促進を目的として，ユーザが作成した文章を第三者に提供する試みが数多く行われている．データの提供にはプライバシを保護するために匿名化が必要となるが，匿名化は攻撃者の知識を前提としているため，匿名化後も攻撃者の知識に応じて機密情報が漏洩する危険がある．そこでオリジナルのデータを提供する代わりに，差分プライバシを満足する合成データを生成する方法が提案されている．差分プライバシは敵対者の知識を前提としないため，匿名化よりも柔軟にプライバシ保護が可能である．しかし，差分プライバシを満足するために生成モデルの勾配に大きなノイズを加えると，生成した疑似データの有用性が損失してしまう．この差分プライバシでは重複を含むデータには比較的少量のノイズで差分プライバシを満足できるという性質があるため，データの有用性低下を防ぐ生成を行うには，ノイズを付加する前に重複を作成する必要がある．そこで本研究では，学習前に概念関係を保持する固有表現を一般化することによってテキスト内に重複を作り出し，ノイズ付加量を低減して勾配の劣化を抑制する．これにより疑似テキストの提供が可能な差分プライベート生成モデルを構築し，変換された疑似テキストの提供によってテキストのプライバシ保護をしつつデータ利活用を促進する．

キーワード　差分プライバシ，疑似データ生成，有用性低下抑制，ノイズ付加量低減

1 はじめに

ブログやウェブページ，消費者生成メディア (CGM) のようなユーザ記述テキストの普及に伴い，新規の研究分野やマーケティングでのデータ利活用を目的として第三者である機関や組織に対してテキストを提供する試みが行われている [4]．しかしこうしたテキストには個人の特定につながるセンシティブ情報が含まれることがあるため，第三者にテキストを提供する際には，個人を識別できないような匿名加工が必要である．

プライバシ保護の手法として，抑制や一般化による匿名化が挙げられるが，これらの匿名化は攻撃者の知識を過程する必要があり，仮定外の場合には対応することができない [15]．したがって，プライバシ保護の観点からは厳密な保護手法ではない．匿名化以外の厳密なプライバシ保護手法として，差分プライバシが挙げられる．差分プライバシとは，誰かが分析結果を閲覧しても個人情報に対して同じ推論しかできないことを数学的に保証するものであり，攻撃者の知識を仮定する必要がない．

こうした背景から差分プライバシを適用したデータ提供が注目を集めており，昨今では深層学習によるデータ生成を行い，提供するという方法が提案されている．深層学習では，学習を繰り返す際に確率的勾配降下法 (SGD) によってパラメータを更新するが，学習の際にこの勾配において個人の識別につながる情報が含まれていると，生成されるデータからも個人の識別が可能となってしまう可能性がある．そこで学習によって勾配を更新する際にノイズを加えることで，個人の特定につながる情報が含まれない最適化を行い，個人が特定できない疑似データを生成可能である．生成した疑似データのみを第三者へ提供することで，元のデータを一切公開しない秘匿なデータ提供を行うことができる [1, 13]．

一方テキストは数値データとは異なり，人名や地名，組織名等の固有表現が含まれることで一意な単語が数多く含まれている．したがってそのまま学習させてしまうと生成モデル構築の際に差分プライバシの性質から勾配に対して大きなノイズの付加が必要となってしまい，有用なテキストを生成するモデルの構築ができなくなってしまう．これに対して，差分プライバシにはレコードが一つ抜けた場合とそうでない場合の出力が区別しづらいことから，重複が多数含まれるデータではノイズ付加量が低減するという性質がある．

そこで本研究では，重複を含むデータに対しては比較的小さいノイズを加えるだけで差分プライバシを満足することができるという差分プライバシの性質を活用して，学習を行う前に固有表現の一般化によって重複を複数作成した．これによって勾配に加えるノイズ付加量の肥大化を抑えた．学習時に勾配に対してノイズを加えることで差分プライバシを満足するテキスト生成モデルを構築し，個人の識別が不可な疑似テキストを生成を行う．

本論文の構成は以下の通りである．まず 2 節では，一般化の関連研究の概要とその問題点について説明を行う．3 節では，保護できるプライバシを明確にするために，本研究で対象とする攻撃モデルについての説明を行う．4 節では，既存の問題点に対応するために行った提案の内容について説明を行う．5 節では，提案手法の堅牢性と有用性を検証するために行った評価実験の内容と結果に対する考察について説明を行う．6 節では，

本研究のまとめと今後の課題について説明を行う．

2 関 連 研 究

本節では基礎事項および関連研究として，差分プライバシの基礎事項と擬似データの生成に関して説明を行う．

2.1 (ϵ, δ)-差分プライバシ

差分プライバシとは，データベースに含まれる個人のデータを保護するための数学的なプライバシ保護指標である．以下の定義 1 は (ϵ, δ)-差分プライバシの定義について説明したものである．

［定義 1］ たかだか 1 レコードしか異ならない隣接関係にあるデータベース D と D' において，ランダム化機構 $\mathcal{M} : \mathcal{D} \mapsto \mathcal{R}$ は (ϵ, δ)-差分プライバシを全ての出力集合 $\mathcal{S} \subseteq Range(\mathcal{M})$ に対して満足する．

$$P[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\epsilon) \cdot P[\mathcal{M}(D') \in \mathcal{S}] + \delta \qquad (1)$$

式 (1) では,(ϵ, δ)-差分プライバシは ϵ-差分プライバシを δ の確率だけ許容すると解釈される．この (ϵ, δ)-差分プライバシは多くの研究で用いられており [2, 5, 12], 本研究もこの (ϵ, δ)-差分プライバシを採用する．また差分プライバシの原理から，プライバシの損失はランダム化機構の出力 o が与えられたときに入力された任意のデータベース D が何であったかの推定ができるかどうかによって評価することができる．

$$l\left(o; \mathcal{M}, D, D'\right) \triangleq \ln \left| \frac{\Pr[\mathcal{M}(D) = o]}{\Pr\left[\mathcal{M}\left(D'\right) = o\right]} \right| \qquad (2)$$

式 (2) の $l(o; M, D, D')$ はデータベース D を用いた場合とデータベース D' を用いた場合の識別可能性を表している．この確率が大きいほど攻撃者は識別が可能なため，プライバシロスが大きい．

2.2 差分プライベートな生成モデル

従来のプライバシ保護技術には k-匿名性 [16] や l-多様性 [9] を満足するように加工する匿名化が用いられてきたが，匿名化では攻撃者の知識を仮定しなければならない．そのため匿名化後であっても攻撃者の保持する事前知識に応じて個人のプライバシが侵害される危険がある．そこで擬似テキストを生成する深層学習モデルを用いて元テキストの学習を行い，深層学習モデルやモデルから生成される擬似データのみを提供することで，プライバシを保護する手法が提案されている [1, 13]．深層学習モデルでは学習の際に確率的勾配降下法によって勾配を更新するが，この勾配において個人のプライバシを侵害する可能性のあるデータを保持してしまう．近年では深層学習モデルに対する攻撃も研究されていることから，こうした勾配自体もセキュアにしてセンシティブ情報の漏洩を防ぐ必要性がある．

そこで学習の際に勾配に対してノイズを加えることで，差分プライベートなモデルを構築する手法が提案されている [1]．この手法では，確率的勾配降下法によって勾配を更新する際にガウシアンノイズを付加しており，これによって (ϵ, δ)-差分プライバシを満足する．全てのサンプルに対してガウシアンノイズを加えることで，どのサンプルもそれ一つで出力結果に影響を与えることがなくなる．一方，確率勾配降下法は入力される値に一意な要素が多く含まれると，最適化する課程で学習が適切に収束しない．そのため正しく最適化させるには，特定の生成タスクに特化したロバストなモデルを構築するか，学習前にこうした一意な要素への対処が必要となる．本稿で対象とする個人が自由記述可能なテキストでは，固有表現のように一意な要素が大量に含まれることからこの問題の解決は必須である．

生成モデルは現在でも様々なものが提案されており，今後もより高精度のものが提案されることが想定されるため，本研究ではロバストなモデルではなく，データ前処理と最適化方式に関する提案を行う．

2.3 ノイズ付加量低減

多量のノイズ量を付加してしまうと，データの有用性を著しく損なう．そこで，ノイズ付加量をへらすための試みとして差分プライバシを満足するためのノイズを付加する前に k-匿名性を先に満足する手法を提案している．この研究では，ノイズを加える前に k-匿名性を満足するようにデータを一般化することで重複を複数作成している．重複が多数ふくまれるデータではレコードが一つ抜けた場合とそうでない場合の出力が区別しづらいことから，差分プライバシを満足しやすい．そのため一意な値が重複となるように一般化し，付加するノイズ量を減らしている [14]．

テキストにおいて含まれる固有表現に対しても，同様にノイズを加える前に k-匿名性を満足するように加工することで，加えるノイズ量をへらすことができると考えられる．しかし数値データと異なり，テキストでは一般化可能な表現とそうでない表現が混合している．そのため，まず一般化可能な単語をテキスト内部から抽出する必要がある．我々は，テキストから一般化可能な固有表現として地名と組織名を抽出して，地理情報システムを介して k-匿名性を満足する手法を提案し，実際の地理情報を考慮した正確な一般化と表記ゆれへの対応を実現した [17]．そこで本稿ではこの手法を一部適用し，テキストに含まれる固有表現を抽出および一般化することで k-匿名性を満足させる．こうして一意な固有表現が含まれない形に加工した状態で学習を行い，勾配に対してノイズを加えつつ生成モデルの学習を行う．

3 想定シナリオ

この節では，提案手法の理解を促進するために本研究で対象とする攻撃シナリオの説明を行う．

本研究では，定められたプロトコルに従って取得した情報から最大限に他人の情報を盗ろうとする攻撃モデルを仮定する．そのため，この攻撃モデルでは参加する二者がどちらもプロトコルの仕様から逸脱した行為を行わないが，プロトコルの仕様内で得られた中間結果や出力等から非公開情報を最大限に窃取しようとする．図 1 は匿名化加工した後のテキストが分析者兼攻撃者に対して公開されたことで，テキストを記述したユーザ

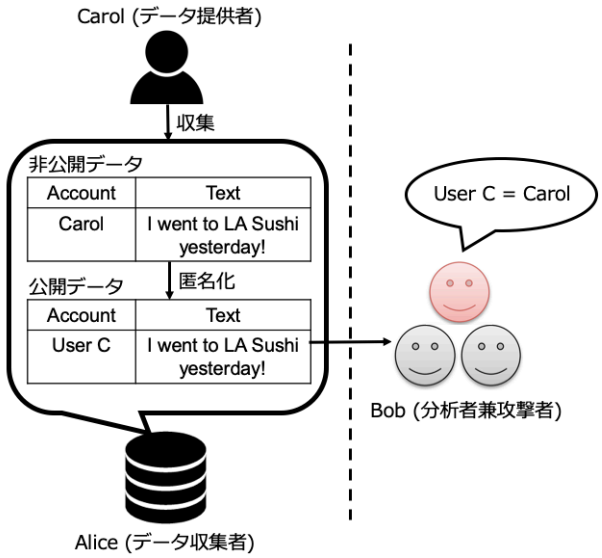

図 1　本研究の想定シナリオ．データ収集を行ったサービスプロバイダが匿名化した後に分析者兼攻撃者にデータを渡すことで，混じっている攻撃者はどのユーザーがどのテキストを書いたのか識別しようとする．

が一意に識別されてしまい，どのユーザが記述した文章なのかを推論される例を表している．このようにテキストを記述したユーザが一意に識別されてしまうことで，記述したテキストと合わせてユーザの属性情報まで推論されてしまう場合がある．例にあげると，記述したテキストに出先の情報が含まれていた場合，過去に記述したテキストからユーザの住所や所属組織等のあらゆる個人情報が推論可能となってしまう場合がある．したがって提供したテキストから，記述した著者が特定されてしまうことはプライバシの侵害につながるため，このようなシナリオを想定してテキストの匿名加工を行う必要がある．

4　提 案 手 法

本節では，我々の提案手法について説明を行う．図 2 は提案手法の全体フローを表している．まず共通の集合を作成するために固有表現を一般化した後，ガウシアンノイズを加えながら生成モデルの学習を行う．これによって差分プライバシを満たす生成モデルを構築し，プライバシの保護されたテキストへの変換を試みる．

本研究の提案は次の 3 つに分けられる．

（1）　元テキストの一般化によるノイズ付加量の低減
（2）　差分プライベート生成モデルの構築
（3）　元テキストの分布に基づく擬似テキストの生成

図 2 は本研究の提案するアルゴリズムのアウトラインである．テキストにおいて含まれる固有表現に対しても，同様にノイズを加える前に k-匿名性を満足するように加工することで，加えるノイズ量をへらすことができると考えられる．我々は実際の地理情報を考慮して，テキストにおける地名や組織名を一般化して k-匿名性を満足する手法を提案した．そこで本稿ではこの手法を用いて固有表現を一般化し，k-匿名性を満足させる．こうしてユニークな固有表現が含まれない形に加工した状態で勾配に対してノイズを加えつつ生成モデルの学習を行うことで，差分プライベートなテキスト生成モデルを構築する．なお，今回の擬似テキスト生成では評価実験において二値分類を想定しており，生成モデルの汎用的な性能ではなく極性判定の精度を重視してラベルごとに最適化された生成モデルを構築した．

4.1　固有表現の一般化

2 節で説明したように，重複を作り出すことで差分プライバシを満足するために加えるノイズ量を減らし，有用性の低下を防ぐことが可能になると考えられる．数値データや画像データの場合，一意なレコードや画素を事前に一般化することで，ノイズ付加量を低減させることが可能である．しかしテキストの場合，数値データや画像データのように一般化をするにはまず一般化する対象の選定と，一般化することが可能な上位概念や下位概念をもつ語を選定する必要がある．過去の研究によると，人名と組織名，地名は固有表現のなかでも個人の特定に大きく寄与することから，一般化対象とされることが多い [10]．固有表現とは，人名や地名，組織名といった固有名詞を体系的にまとめたものの総称である．

そこで本研究では，これら三種類の固有表現を事前に一般化する．2 節で述べたとおり，ナレッジグラフを用いて地名と組織名を一般化することが可能である．しかし，この研究では人名に対してはアプローチしておらず，本研究でそのまま適用すると人名については匿名化されずに残してしまう．そこで本研究が人名をどのように一般化するのかについて説明を行う．図は本研究で人名を一般化する手順について説明している．固有表現抽出器によって抽出した人名に対して一般化を行い，k-anonymity を満足するまで一般化していく．またイニシャルのみの状態でも k-anonymity を満足することができない場合，アスタリスクで置き換えることで対応する．なお本研究では，あらゆるテキストにおいて高精度の固有表現が抽出可能な BERT [7] を用いて固有表現抽出を行っている．

4.2　差分プライベート生成モデルの学習

生成モデルには RNN [11] や LSTM [6] 等の深層学習を用いたものが数多く提案されており，これらの学習にはセンシティブな情報を含むデータを用いて行う．このセンシティブな情報を含むデータを学習して構築したモデルの勾配は，保護されなければならないプライバシに関する情報を含む．学習済みのモデルを公開して第三者に提供することはしばしば行われているが，近年では出力値や勾配から学習データを推測可能であるというメンバシップ推論攻撃が脅威として挙げられている．したがって，そのまま出力値や勾配を提供してしまうのは危険である．

そこで本研究では，本研究も差分プライバシを満足する勾配をもつ生成モデルによって変換したテキストを提供する手法を提案する．Algorithm 1 は提案する最適化法である．一つのテキストは単語のシーケンスであるため，単語同士には依存関係が必ず含まれる．そのため，この依存関係を長期的に学習可能

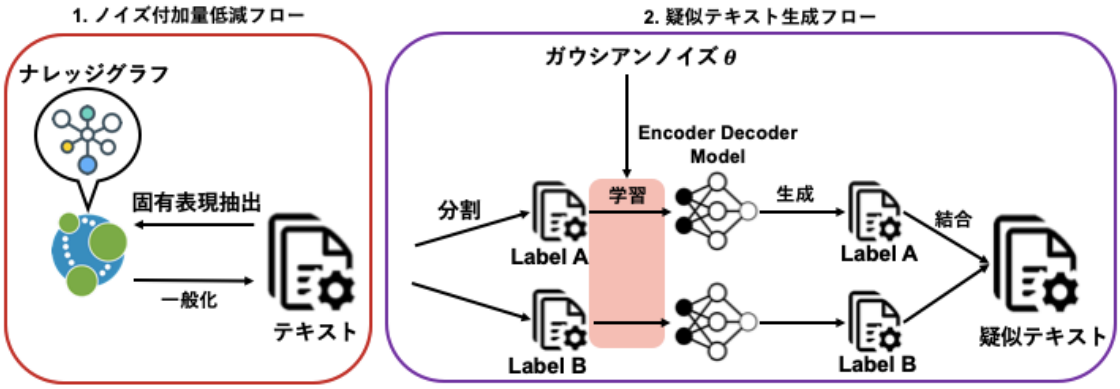

図 2　提案手法の全体的なワークフロー．入力されたユーザ記述テキストは Knowledge Graph [17] によって一般化され，ラベルに基づいて分離される．各ラベルに基づいて生成モデルを構築し擬似テキストを生成し，各ラベルの擬似テキストを結合する．

Algorithm 1 Adam による差分プライベートな最適化

Input: $\mathcal{X}:\{x_1,x_2,\cdots x_n\},\mathcal{J}(\cdot),\eta,\sigma,\mathcal{L},\mathcal{C}$
Output: θ_T

Initialize θ_0, m_0, v_0
for $t \in [T]$ **do**
  select random sample L_t where sampling probability $\frac{L}{N}$
  for each $i \in L_t$ **do**
    $g_t(x_i) \leftarrow \nabla_{\theta_t}\,\mathcal{J}(\theta_t, x_i)$
  end for
  $\overline{g}_t(x_i) \leftarrow \max\,(1, \frac{\|g_t(x_i)\|_2}{C})$
  $\tilde{g}_t \leftarrow \frac{1}{L}\sum_i(\overline{g}_t(x_i)) + \mathcal{N}(0,\sigma^2C^2I)$
  $m_{t+1} = \beta_1 m_t + (1-\beta_1)\,\tilde{g}_t$
  $v_{t+1} = \beta_2 v_t + (1-\beta_2)\,\tilde{g}_t^2$
  $b_{t+1} = \frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}}$
  $\theta_{t+1} = \theta_t - \alpha_t \frac{m_{t+1}}{\sqrt{v_{t+1}+\phi}} b_{t+1}$
end for

な LSTM を基調としたエンコーダデコーダモデルとした．なお，Adam に含まれるパラメータは論文内で推奨されている $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\phi = 10^{-8}$ を用いている．

4.3　差分プライベートなテキストの変換

提案生成モデルでテキストをプライバシが保護されるように変換するには，どのような単語をもとに生成するか，また生成するテキストの長さはどのくらいかをパラメータとして与える必要がある．この際に適当な単語を入力してしまうと，もとのテキスト集合とは全く異なる擬似テキスト集合となってしまう可能性がある．したがって，もとのテキスト集合における単語の出現確率に基づいて変換を行う必要がある．

$$P(T_w|p) = \prod_{w=1}^{n} p^{T_w}(1-p)^{1-T_w} \qquad (3)$$

ここで，$T_1, T_2, \cdots T_n$ は二項分布 B(1,p) に従うことから，$P(X_i|p) =$ と変形することができ，A の総乗として表すことができる．この出現確率に基づいて，我々はオリジナルのテキストから確率的ランダムサンプリング (確率抽出法) によってトークンを抽出する．

Algorithm 2 確率的ランダムサンプリングによる変換

Input: $\mathcal{X}:\{x_1,x_2,\cdots x_n\},\mathcal{G}_\theta(\cdot),l_i$
Output: $\mathcal{X}^{dp}$

compute $P(T_w|p)$
for $w \in [d]$ **do**
  select token T_w where sampling probability $P(T_w|p)$
  generate differentially private text x_i^{dp} by $\mathcal{G}_\theta(T_w, l_i)$
  add x_i^{dp} differentially private text set $\mathcal{X}^{dp}$
end for
return $\mathcal{X}^{dp}$

5　評 価 実 験

この節では，有用性損失の抑制についてと，元のテキストと生成したテキストにおける非可逆性に関する評価実験について説明を行う．図 3 は，評価実験のアウトラインを表している．

5.1　有用性損失の抑制

差分プライベートな生成モデルによってテキストを生成しても，元のテキストと比較するとデータセットに与えられているタスクに用いることができない可能性がある．そこで生成前のテキストを用いた場合のタスク結果と比較して，どの程度精度を低下させてしまうのか評価を行った．なお評価には，Large Movie Review Dataset [8] を用いた極性判定を行う．

この際の差分プライベート生成モデルにおけるのハイパパラメタには，学習のモデルの epoch 数を 500，LSTM におけるはドロップアウト比率は 0.2，中間層の活性化関数には，出力層の活性化関数には Softmax 関数を用いている．そしてモデルコンパイルの損失関数には，特定の単語のみが生成される局所最適化を抑制するためにクロスエントロピ，最適化関数には Adam を用いた．

分類に使用したモデルは多項ナイーブベイズ，サポートベクトルマシン，多項ロジスティック回帰，決定木 (C4.5) の四種類である．これらを用いて分類した結果が以下の表である．

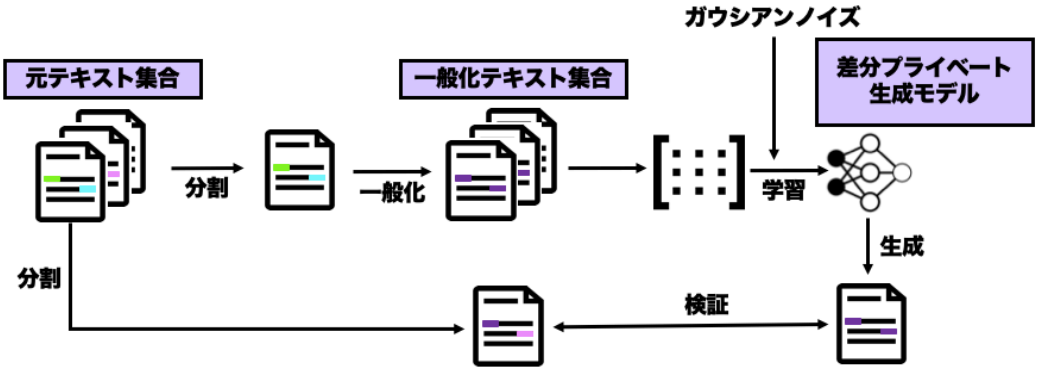

図 3　提案手法とベンチマークの比較に関する実験概要

表 1　Utility Loss：Results of Polarity Classification

	Original Text		DP E-D		**DPvG E-D(Ours)**	
	POS	NEG	POS	NEG	POS	NEG
MLR	0.582	0.681	0.426	0.622	0.510	0.648
SVM-RBF	0.642	0.723	0.577	0.653	0.609	0.680
MNB	0.640	0.755	0.595	0.661	0.614	0.691
DT-C4.5	0.609	0.739	0.536	0.639	0.570	0.673

実験の結果，一般化処理なしのエンコーダデコーダモデルによるテキストを用いた場合よりも，提案手法の一般化処理後エンコーダデコーダモデルによるテキストを用いたほうが極性判定における F1-Score が通常のエンコーダデコーダモデルを用いた場合に近いため，より有用性損失を抑制したテキストを生成できたことがわかる．これには固有表現が一般化されたことで従来のテキストと比べて低頻度の単語も生成され，変換した後のバリアンスが向上したことが寄与したためである．

5.2　変換した擬似テキストの評価

変換したテキストから元のテキストが推論可能な場合，プライバシを保護したとは言えないため，元のテキストとの変換した擬似テキストは非可逆である必要がある．また差分プライバシ適用後に出力結果に一定の無作為性を確認できない場合，同様にプライバシを保護したとはいえない．したがってオリジナルテキストと変換したテキストのコサイン類似度行列の差異，およびプライバシロスの変動を評価する．

コサイン類似度行列は，元の文書 d とプライバシが保護されるように変換した文書 d^{dp} のコサイン類似度を要素にもつ行列である．このとき文書 d と文書 d^{dp} はそれぞれ各単語の発生確率による重みを下げるために，Residual-IDF [3] による重み付けを行ってコサイン類似度の計算している．こうした文書のベクトル化には，単語の出現頻度である Term Frequency(TF) と単語の逆文書頻度である Inverse Document Frequency(IDF) という二つの情報から単語の重要度を算出する TF-IDF が一般的だが，TF-IDF では文書ごとの単語数の差による影響がでてしまう．そこで本研究では，ポアソン分布による重み付けを行い，特徴的な単語の重みを下げることなく文書間で共通して現れる単語の重みのみを下げることが可能な Residual-IDF でベクトル化する．

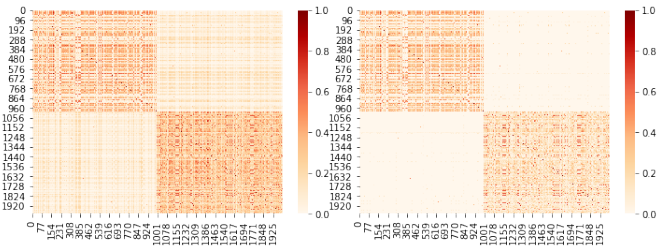

図 4　左: 差分プライベート生成モデルを用いて学習前に一般化を行わずにコサイン類似度を表現したヒートマップ．右: 差分プライベート生成モデルを用いて学習前に一般化を行った場合のコサイン類似度のヒートマップ

ガウシアンノイズの付加量を調整した場合とそうでない場合のコサイン類似度行列において差が生じており，ガウシアンノイズを調整した場合のほうが元の文章と類似していないことがわかる．これは，ガウシアンノイズの付加量を調整して学習するために，事前に固有表現を一般化したことで，元の文章にふくまれていた固有表現等が変換されないことで，コサイン類似度が通常の差分プライベード生成モデルよりも小さくなったと考えられる．

また，加えたノイズ量が減ったことでどの程度のプライバシロスが生じているのかを計測するために，式 2 に従っては各単語を入力とした際の出力される単語の最大確率とした場合のプライバシロスを算出した．図 4 は privacy budget と privacy loss の関係を表すグラフである．DP E-D は提案手法を適用しない場合の差分プライベートなエンコーダデコーダモデルであり，DPvG E-D は提案手法によるエンコーダデコーダモデルである．プライバシロスは 0.8 よりも大きい場合に有意な差があった．実験の結果，0.8 以上では DP E-D の方が DPvG E-D よりもプライバシロスは抑えられているという結果になった．これは，DP E-D の方が加えられたノイズ量が多く，出力に偏りが生じているためだと考えられる．DPvG E-D では加えたノイズ量が DP E-D と比較すると少なかったことで，プライバシロスが大きくなってしまったと考えられる．なお 0.8 未満では局所最適解に陥ってしまい，多様な単語生成ができなかったため，プライバシロスが生じなかった．この指標は言い換えると無作為性を評価するものであり，提案手法である DPvG

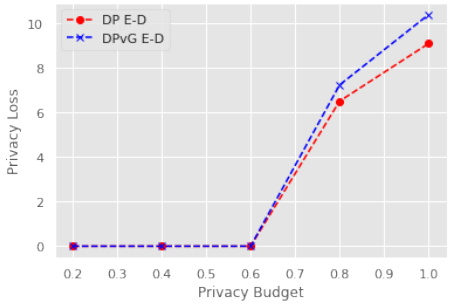

図 5　プライバシバジェットとプライバシロスの関係を示す折れ線グラフであり，DP E-D は提案手法を用いない場合の差分プライベートエンコーダデコーダモデル，DPvG E-D は提案手法を用いた場合のエンコーダデコーダモデル．

E-D の方が多様な単語生成が可能となったとも言える．

5.3 議　　論

これらの結果から，感情分析というタスクを目的とした場合のデータの有用性を低下させることなく，データを堅牢にすることができたと考えられる．また今回の提案では極性判定に特有の提案ではないため，機械翻訳や文書要約のような他の自然言語処理タスクにおいても使用可能だと考えられる．

一方，本研究の限界としては地名や組織名，人名が大きく影響するタスクへの応用が困難であることが挙げられる．付加するガウシアンノイズを低減するために事前に固有表現を一般化したことで，これらの固有表現が直接与えられているタスクに影響する場合は，テキストの有用性を保証できないと考えられる．

6 おわりに

本稿では，テキストを直接提供するのではなく，テキストを生成する差分プライベートな生成モデルを構築し，生成されたテキストのみを提供することでプライバシを保護しつつ有用なデータを提供する枠組みを提案した．評価実験の結果，通常の差分プライベード生成モデルよりも有用なテキストの生成が可能となり，かつプライバシも保護された．

今後の課題は三点ある．一つ目は昨今話題の BERT のような Transfomer モデルを用いた生成を検討すること，二つ目は他の固有表現も一般化することで加えるガウシアンノイズを同様に低減させること，最後は今後ラベルが増えた場合に対応するためにより Conditional なモデルに発展させることである．

謝　　辞

本研究は JSPS 科研費 JP18H03234 および ICS-CoE 中核人材育成プログラムの助成を受けて遂行されたものである．また本稿執筆にあたりアドバイスをいただいた株式会社サイボウズラボの中谷秀洋氏に感謝の意を記す．

文　　献

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] A. Beimel, K. Nissim, and U. Stemmer. Private Learning and Sanitization: Pure vs. Approximate Differential Privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 363–378. Springer, 2013.

[3] K. Church and W. Gale. Inverse Document Frequency (idf): A Measure of Deviations from Poisson. In *Natural language processing using very large corpora*, pages 283–295. Springer, 1999.

[4] B. C. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):1–53, 2010.

[5] Q. Geng and P. Viswanath. Optimal Noise Adding Mechanisms for Approximate Differential Privacy. *IEEE Transactions on Information Theory*, 62(2):952–969, 2015.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] J. Li, A. Sun, J. Han, and C. Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[8] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3–es, 2007.

[10] N. Mamede, J. Baptista, and F. Dias. Automated Anonymization of Text Documents. In *2016 IEEE Congress on Evolutionary Computation*, pages 1287–1294. IEEE, 2016.

[11] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur. Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing*, pages 5528–5531. IEEE, 2011.

[12] A. Nikolov, K. Talwar, and L. Zhang. The Geometry of Differential Privacy: The Sparse and Approximate Cases. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 351–360, 2013.

[13] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim. Data Synthesis based on Generative Adversarial Networks. *Proceedings of the VLDB Endowment*, 11(10), 2018.

[14] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. Enhancing Data Utility in Differential Privacy via Microaggregation-Based *k*-Anonymity. *The VLDB Journal*, 23(5):771–794, 2014.

[15] L. Sweeney. Achieving k-anonymity Privacy Protection using Generalization and Suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[16] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[17] S. Taisho, T. Yuzo, and K. Youki. Anonymizing Location Information in Unstructured Text Using Knowledge Graph. *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications and Services*, 2020:163–167, 2020.