位置推定とその実現可能性を考慮した新しい匿名化の提案

田口 勝弥[1]　若宮 翔子[1]　荒牧 英治[1]

概要：個人情報の取り扱いについて関心が高まるにつれ，匿名化に関する需要が高まっている．本研究では，位置情報の匿名化に焦点を当てる．位置情報の匿名化については，これまで，GPS 情報や住所などのテキスト表現をマスクする手法が多かった．しかし，思わぬ語や句の組み合わせで位置情報が判明してしまう場合もあり，複雑な問題となっている．本研究では，「位置情報の特定」について，位置（座標やエリア）推定，もしくは，位置が分かりそうかどうか（本稿では位置推定の実現可能性と呼ぶ）という 2 つの観点から，多段階の匿名化レベル，ならびに，任意のレベルまで自動匿名化する手法を提案する．まず，位置情報付き発言から作成した位置を推定する分類器を構築し，次に発言について位置推定の実現可能性を持つかどうかタグ付けを行った．さらに，これらを用いて，機械では位置を推定できる／推定できないが，人間には位置が分かりそう／分かりそうにない，という複数のレベルに発言を分類する．どのレベルを位置情報の匿名化とみなすかは用途に依存する．本研究は，位置を「特定できる」ことと「特定できそう」であることの両方を考慮した新しい匿名化研究である．

キーワード：匿名化，位置推定，SNS，Twitter，自然言語処理

Novel De-identification using Location Estimation and Its Feasibility

Katsuya Taguchi[1]　Shoko Wakamiya[1]　Eiji Aramaki[1]

Abstract: Nowadays, personal information has drawn much attention, requiring the advanced technology on de-identification. This study focuses on location information. The conventional approaches remove the GPS information or expressions such as addresses. However, there can be a complicated case where location information can be estimated with unexpected combinations of non-address words. To deal with this phenomenon, this study proposes two types of de-identification level. One point is to estimate location, and the other is to estimate whether location can be estimated or not (what we call "feasibility of location estimating "). To realize both levels of de-identification, first, we make with geo-tagged texts a classifier which estimates locations. We, next, tagged texts with feasibility of location estimating. By using the classifier and newly tagged texts, we classify texts with these classifiers into four levels. We believe our novel concepts on de-identification are essential for various practical applications.

Keywords: De-identification, Location estimation, SNS, Twitter, Natural Language Processing

1. はじめに

近年，プライバシー保護への要請が高まっていることから，個人を特定する情報を削除する技術である自動匿名化の研究が活発になされている．多くの自動匿名化研究で取られる方法では，個人を特定する情報の大部分が人名，組織名，電話番号，ID，住所などの固有名であることから，自然言語処理の固有表現認識を拡張することでこれらを削除し匿名化している．しかし，実際には固有名以外から住所が判明する場合も多い．例えば，「川床での食事は絶品だね」という発言において，固有名は存在しないが，「川床」が存在する場所は限られており，例えば，京都市内に限定すると 200 mほどの範囲で特定可能となる．ソーシャル・ネットワーキング・サービス（以下，SNS）では，上記のような個人の特定に利用できる情報を，本人が意図せ

[1] 奈良先端科学技術大学院大学

ず発信してしまう事例が多く報告されており，社会問題となりつつある．また，人間ならば分からないような位置情報の漏洩でも，機械であれば位置を推定可能である場合があり，今後機械学習の精度が向上することを考えれば，位置情報を意図せず漏洩してしまうことはますます問題となるであろう．

本研究では，代表的な SNS である Twitter への発言から，位置情報をマスクする自動匿名化を扱う．本研究で扱う自動匿名化は次の 2 点が新しい．

(1) 対象の固有名だけでなく，あらゆる単語の組み合わせを考慮している点．
(2) 匿名化された状態に複数のレベルを想定している点．

（2）に関しては，特に，機械に位置が推定できるか／人間に位置が推定できるかの組み合わせによって匿名化のレベルを定義する．さらに，この（2）を実現するため，人間，機械のそれぞれに対応した位置推定の分類器を構築し，それぞれ位置が不明となるまで単語を削除することで自動匿名化を行う．

本稿の構成は以下である：まず，Twitter に投稿された位置情報付きの発言を用い，発言の位置を推定する分類器を構築する（4 章）．次に，人間が匿名化した状態にした発言について，分類器が位置を推定できるかどうかを調査する（5 章）．さらに，発言から，人間がその発言位置を推定できそうかどうか（本稿では**位置推定の実現可能性**と呼ぶ）を学習し，位置推定可能としたテキストとの比較を行う（6 章）．最後に，これらの知見とデータをもとに，単語の組み合わせを考慮した匿名化手法を提案する（7 章）．

2. 関連研究

2.1 位置推定

位置推定に関する先行研究を以下の表 1 にまとめる．

まず，位置推定に用いる確率の種類によって 2 種類のモデルが存在する．1 つは word-centric と呼ばれるモデルであり，単語の集合 w に対し，位置ラベル l を出力する確率 $p(l|w)$ を求める．もう 1 つは location-centric と呼ばれるモデルであり，位置ラベル l が文書 d を出力する確率 $p(l|w)$ を求める．本稿では word-centric モデルを用いて分類器の構築を行う．また，SNS に投稿されるテキストから推定される位置についても user home locations，tweet locations，mentioned locations の 3 種類のタイプが存在する．user home locations は投稿者の住所などにあたる．tweet locations は投稿された位置にあたる．mentioned locations はその投稿によって言及されている位置にあたる．本稿では tweet locations をマスクする匿名化を対象とする．最後に，位置推定に用いる材料についても Twitter network，tweet content，tweet context の 3 種類が存在する．Twitter network とはユーザー間のフォロー，フォロワーの関係のことである．tweet content とは投稿されるテキストの内容のことである．tweet context とは投稿されるテキストに付随するジオタグや時刻に関する情報のことである．本稿では tweet content のみから位置情報を推定する．

文献 [6] は，tweet content から word-centric モデルを用いて tweet locations を推定した研究であるが，地域ごとの特徴語のリストを持つように分類器を学習させる．しかし実際には，1 章でも述べたように思わぬ単語および単語の組み合わせによりその発言位置が特定される場合を想定することが可能である．このような場合を想定し，本稿では，形態素の組み合わせによって匿名化する手法を提案する．

2.2 匿名化

テキストの匿名化については，医療分野での患者データ匿名化など研究が活発に研究されている．しかし SNS での発言をその発言位置に関して匿名化する場合，患者データの場合のように人名や電話番号をマスクするだけでは不十分である．また，医療分野では HIPAA[*1]において匿名化の明確な評価基準が設けられているが，SNS へ投稿されるテキストの匿名化については HIPAA にあたるものが存在しない．そのため，テキストがその発言位置について匿名化されていると決定するための基準を新たに設ける必要がある．この点を考慮し，本研究では人手で匿名化の実験を行なうことで基準の設定を行う．

3. データセット

本章では，以後の各章で用いるデータ，また対象とするエリアの区画について述べる．

3.1 テキストデータ

本節では，分類器の構築，および実験に用いるデータについて述べる．

全データとして，京都市中心部において投稿された位置情報つきツイート（以降，発言と呼ぶ）298,711 件を用いた．本稿では，京都市中心部を北緯 34.93 度から北緯 35.12 度，東経 135.67 度から東経 135.83 度の範囲と定義した．この範囲には繁華街である阪急河原町駅周辺や，東海道新幹線の停車駅である京都駅，観光名所として知られる二条城や伏見稲荷大社，清水寺などが含まれるため，多様な特徴を持つ位置からデータが得られる．

3.2 対象エリア

前節で定義した京都市中心部を計 200 エリアとなるように南北に 20，東西に 10 等分割した（図 1）．この分割の仕方に関しては，京都有数の繁華街として知られる阪急河

[*1] 1996 年に米国で制定された，個人の医療データのプライバシーを守りつつ，データを役立てることを目的とした法律 [12]（https://www.hhs.gov/hipaa/index.html）

表 1　位置推定の関連研究

		推定対象		
		Home locations	Tweet locations	Mentioned locations
推定材料	Twitter network	Kong et al. [1]	Sadilek et al. [2]	Hua et al. [3]
	Tweet content	Yamaguchi et al. [4] (word-centric) Cha et al. [5] (location-centric)	Flatow et al. [6] (word-centric) Kinsella et al. [7] (location-centric)	LI et al. [8]
	Tweet context	Efstathiades et al. [9]	Dredze et al. [10]	Fang et al. [11]

図 1　本稿で対象とする京都市中心部の 200 エリア

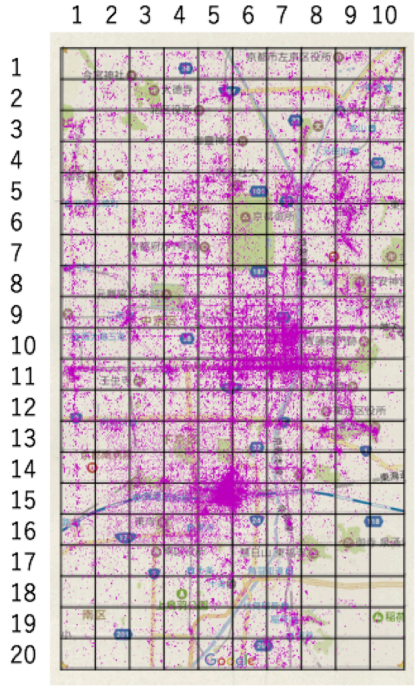

図 2　京都市中心部における発言の地理的分布

原町駅周辺において，連続する 2 つの駅（阪急河原町駅と阪急烏丸駅）を異なるエリアに分布させており，妥当な区画であると考えられる．このように区切られたエリアに対し，元の 298,711 件の発言を分布させた（図 2）ところ，最も発言の多いエリアは京都駅が位置するエリア (15, 5) で，発言数は 39,078 件であった．一方，最も発言の少ないエリアは御寺泉涌寺の南東に位置するエリア (17,10) で，発言数は 2 件であった．

4.　位置推定分類器の構築

本章では，発言位置を推定するための分類器を構築する方法およびその性能について述べる．

4.1　材料

位置推定分類器の構築のために，全データ 298,711 件のうち 60%にあたる 179,227 件を学習データとして用いた．そして，残り 119,484 件をテストデータとして分類器の性能評価を行なった．

4.2　方法と評価

学習データに含まれる各発言を形態素解析器 MeCab[*2] を用いて形態素区切りにし，全 uni-gram と全 bi-gram を合わせたものを素性とする Bag-of-Words としたものをベクトルとして学習を行った．その際，ある形態素が出現したか否かを 1 または 0 で表現するベクトルを用いた．また，「@ユーザ名」などのメンション記号や，URL など学習の際にノイズになると考えられるものを削除した[*3]．正解ラベルは各エリア（全 200 クラス）とし，ロジスティック回帰を用いて学習を行った．

構築した分類器を評価したところ，テストデータ 119,484 件での精度は 47.2%であった．なお，学習データにおいて最も多く発言が分布するエリア (15, 5) を常に出力するような分類器を考えた場合，テストデータに対する精度は 11.6%となった．

5.　予備実験：人手での匿名化

4 章で構築した分類器を用いた場合に，「ある発言が匿名化されている」という状態をどのように定義できるかを調

*2　http://taku910.github.io/mecab/
*3　https://github.com/s/preprocessor

べる必要がある．本章では，匿名化の状態を定義するために行なった予備実験について述べる．

5.1 材料と手順

4 章の分類器の性能評価で用いたテストデータ 119,484 件のうち，500 件の発言について人手による匿名化を行った．

まず，実験参加者が発言を見てその発言位置を可能な限り推定した．その際，参加者は発言中の単語について検索エンジンなどのウェブサービスを用いてよいこととした．次に，実験参加者が発言位置が分からなくなるまで，最小限の形態素を削除することで，人手による匿名化を行なった．なお，本予備実験では，京都市内に土地勘のある 2 名の実験参加者が独立して匿名化を行なった．この結果，以下のような結果が得られた．

(1) 烏丸御池プラザが本チャンやないんか？
~~マザーズハローワーク~~烏丸御池

次に，人手で匿名化を行なった後の発言について，前章で構築した分類器が各エリアに割り当てる所属確率の最大値の平均値を求め，匿名化されていると判定する際の基準値とした．所属確率がこの値を下回っている発言は匿名化されているものとみなす．

5.2 結果と考察

500 件について所属確率の最大値を調べると，匿名化後の発言についてその平均値は 0.37 となった．この結果から，ある発言について分類器が出力するクラス所属確率の最大値が 0.37 を下回った状態を，その発言が匿名化されている状態であると定義する．

しかし，所属確率の最大値について分類器が高い値を出力したものの，被験者が位置を回答しなかったケース，あるいは，その逆のケースが見られた．

このような食い違いが起こることから，匿名化には 2 種類の推定が存在すると言える．すなわち，位置自体の推定を防ぐための匿名化と，位置が推定できそうと思われることを防ぐための匿名化が存在する．したがって，次章では発言から位置が推定できそうかに関してタグ付けを行なったデータを用い，位置推定の実現可能性推定を行う分類器の構築を行う．

6. 位置推定の実現可能性

本章では位置推定の実現可能性推定を行う分類器を構築するための予備実験と実際の分類器の構築について述べる．

6.1 材料と手順

4 章で用いたテストデータ 119,484 件の発言のうち，無作為に抽出した 1,000 件を用いた．大量の実験協力を集めるためにクラウドソーシングでタスクを実施した．発言位置が推定できるかどうかに関して人間が 2 値分類を行う課題を行なった．100 名の実験参加者は各発言に対し，その発言位置が推定可能かどうかを回答した．それぞれの発言に対し，10%以上の参加者が推定可能と回答したものを位置推定の実現可能性のある発言，10%に満たないものを実現可能性のない発言に分類した．

6.2 結果と考察

1,000 件の発言のうち，位置推定の実現可能性のあるものは 246 件であった．対象とした 1,000 件の発言について分類器によって位置が推定できると判定されるかどうか，位置推定の実現可能性があるかを分類した結果の総計データを表 2 に示す．

表 2 分類器と人間の位置推定の実現可能性推定

		分類器	
		推定可能	推定不可能
人間	推定可能	216	30
	推定不可能	258	496

分類器は位置を推定できるものの，位置推定の実現可能性は低くなった発言の例を以下に示す．

(2) 風が強いです (>_<) 今日も明るく元気にお昼の営業開始です！
(3) おはようございます (^^) 今日の日中は雨予報ですね。 気温も 20 ℃まで行かないようです。 今日も明るく元気に！忙しく楽しい一日になるよう頑張ります p(^_^)q

これらの発言は，店の営業目的での定型文を含んでいる．分類器は大量の発言から学習を行うため，単一の店舗からの発言を学習した結果このようなデータが得られたと考えられる．一方人間は多量の発言を読むことはできないため，これらの発言に対して位置の推定は行えない．以下は，位置推定を行う分類器には位置を推定できないものの，位置推定の実現可能性は高くなった発言の一部である．

(4) やっとお昼ご飯。 つばめ
(5) 河合塾の向かいのサブウェイなう！

(4) は実際は固有名詞である「つばめ」が普通名詞でもあるケースである．人手で位置推定の実現可能性をタグ付けしたデータが必要である．(5) は「河合塾」と「サブウェイ」の組み合わせにより位置推定が可能と判断されるケースである．

6.3 位置推定の実現可能性判別を行う分類器の構築

クラウドソーシングで得られたデータ 1,000 件のうち，

900 件を学習データ，100 件をテストデータとして用い，位置推定の実現可能性を判別する分類器を構築した．

4 章と同様に，学習データに含まれる各発言を形態素解析器 MeCab を用いて形態素区切りにし，Bag-of-Words としたものを学習データとした．なお，「@ユーザ名」などのメンション記号や，URL など学習の際にノイズになると考えられるものを削除した．正解ラベルはクラウドソーシングにおいて 10%以上の参加者が特定可能としたものとそうでないものの 2 クラスとし，ロジスティック回帰を用いて学習を行った．テストデータ 100 件の精度（accuracy）は 86%であった．

7. 議論

以上の実験結果から，匿名化には 2 種類の段階があることを示した．したがって匿名化を行う際には匿名化の目的に応じてシステムを使い分ける必要がある．例えば，4 章で構築した分類器を用いて匿名化を行う場合，営業目的の定型文を含むような発言まで匿名化の対象となる．このような発言に関して匿名化の必要がない場合は 6 章で構築した分類器との併用によって必要な発言のみを匿名化すればよい．

本稿で構築した 2 種類の分類器を使用する場合の匿名化手法として，削除する単語の組み合わせと匿名化基準を用いたものを提案する．

7.1 手法

位置推定のための分類器を用いた匿名化アルゴリズムを以下に示す．

Step 0: 削除する形態素数 m を 1 とする．

Step 1: 発言 s から，任意の m 個の形態素 (群) を削除した発言群 (こうして作った発言群の集合を S とする) を作る．

Step 2: Step 1 にて作った発言群について，位置推定分類器によって，京都市内のどの位置にいるか（$a_{1..200}$）の所属確率 の最大値（$prob(s)$）を求める：

$$prob(s \in S) = \max_{a \in a_{1..200}} (s)$$

Step 3: $prob(s)$ が最も低くなる発言を採用し，s_{new} とする

$$s_{new} = \arg\min_{s \in S} prob(s)$$

ここで，$prob(s_{new})$ が閾値 (0.37) を下回る場合，終了する．そうでない場合は，m 形態素の削除による匿名化失敗とみなし，より多数の形態素の削除を試みる（$m++$ として，STEP 1 に戻る）．

7.2 結果と考察

位置推定するための分類器を用いた匿名化の結果の一部を以下に示す．

提案手法による匿名化例

（6）まだまだ ~~新幹線京都駅~~
（7）5 年ぶり 京都 ~~タワー清水~~の舞台から 1 枚 京都の街が一望だね
（8）ランチ (at なか卯 河原町~~五条店~~)
~~折田~~先生なう
（9）京都~~御所~~一般公開中
（10）阪急~~河原町~~なう

* 取り消し線は匿名化された部分を示す．

「新幹線京都駅」は固有名詞であるが，京都市内に「駅」は多数存在するため，「新幹線京都」を削除すれば匿名化できる．この場合，「駅」という地名にカテゴリの関する情報は保持される．同様に，「京都タワー」においては，京都市内に「タワー」は 1 つしか存在しないため，「タワー」部分を削除する必要がある．このように，従来の匿名化のように，固有名詞全体を削除する必要はなく，最小限の操作で匿名化を達成できる．

十分な学習データがあれば，提案手法は有効に働くと考えられるが，以下のような失敗例も存在した．

（11） 撮り飽きもせず 撮り足りもせず ~~京~~都御苑

「京都」と「御苑」だと「御苑」をマスクする方が妥当であると考えられるが，「京都」のみをマスクしており，適切に匿名化できた例とはいえない．

学習データ（n=100..1000) と精度（accuracy）の関係を調べたところ，学習データは十分な量であるとはいえない（図 3）．本章の分類器の構築には人手のラベリングを要するため多量のデータを得ることが難しく，この点も今後の課題とする．

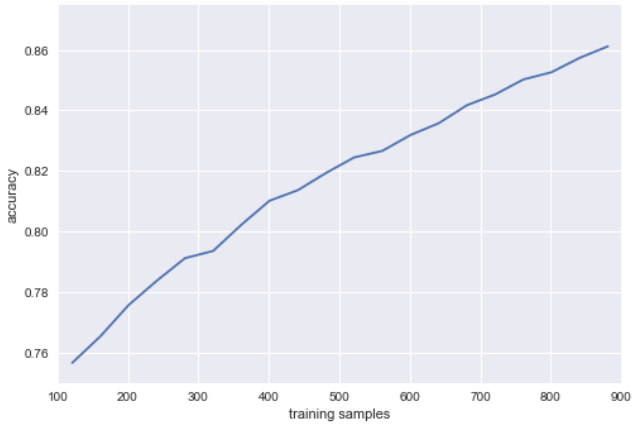

図 3 学習データサイズ (training samples) と精度（accuracy）の関係

8. おわりに

本稿では，SNS 上での発言からその発言位置をマスクするための匿名化を扱った．発言位置に関する推定には，位置情報そのものの推定と，位置情報そのものでなく，推定の実現可能性の 2 種類が存在し，それぞれによって異なった性質がある．前者の場合，店の営業に関する発言など，同一の発言が 1 箇所から投稿されることから，投稿によって位置推定が可能となる．後者の場合，エリアに偏在していない普通名詞が店の名前として用いられている場合のように，人間がこれまでの匿名化に関する研究における諸概念をそのまま利用するだけでは実用において不十分であることを示した．さらに単語の組み合わせによって位置の推定が行えるという仮説に基づいた匿名化アルゴリズムの提案を行い，その結果の一部については目的通りの処理が行えていることを示した．しかし結果の中には，適切に匿名化できていないものも含まれた．

今後の課題としては，文脈の扱いが挙げられる．本研究では各発言を Bag-of-Words 化した．そのため「河合塾の向かいのサブウェイなう！」のように離れた形態素を扱えなかった．今後，Bag-of-Words ではなく，構文解析などの結果も考慮することで解決が見込まれるため，今後の課題とする．

謝辞

本研究の一部は，日本医療研究開発機構研究費 新興・再興感染症に対する革新的医薬品等開発推進研究事業（課題番号：16768699），戦略的情報通信研究開発推進事業（SCOPE）（課題番号：17934316），JSPS 科研費 JP16K16057 および JST ACT-I の支援を受けたものです．

参考文献

[1] L. Kong, Z. Liu, and Y. Huang. Spot: Locating social media users based on social network context. *Proceedings of the VLDB Endowment*, 7(13):pp.1681–1684, 2014.

[2] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proc. 5th International Conference on Web Search and Web Data Mining (WSDM)*, pp. 723–732, 2012.

[3] W. Hua, K. Zheng, and X. Zhou. Microblog entity linking with social temporal context. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1761–1775, 2015.

[4] Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa. Online user location inference exploiting spatiotemporal correlations in social streams. In *Proc. 23rd ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1139–1148, 2014.

[5] M. Cha, Y. Gwon, and H. T. Kung. Twitter geolocation and regional classification via sparse coding. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, pp. 582–585, 2015.

[6] D. Flatow, M. Naaman, K. E. Xie, Y. Volkvich, and Y. Kanza. On the accuracy of hyper-local geotagging of social media content. In *Proc. 8th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 127–136, 2015.

[7] S. Kinsella, V. Murdock, and N. O'Hare. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proc 3rd International CIKM Workshop on Search and Mining User-Generated Contents (SMUC)*, pp. 61–68, 2011.

[8] G. Li, J. Hu, J. Feng, and K.-l. Tan. Effective location identification from microblogs. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pp. 880–891. IEEE, 2014.

[9] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos. Identification of key locations based on online social network activity. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM*, pp. 218–225, 2015.

[10] M. Dredze, M. Osborne, and P. Kambadur. Geolocation for twitter: Timing matters. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1064–1069, 2016.

[11] Y. Fang and M. Chang. Entity linking on microblogs with spatial and temporal signals. *TACL*, 2: pp.259–272, 2014.

[12] 荒牧英治. 医療言語処理（自然言語処理シリーズ 12），コロナ社，2017.