様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　２４　日現在

機関番号：　１４６０３
研究種目：　挑戦的萌芽研究
研究期間：　2013～2015
課題番号：　２５５４００９６

研究課題名（和文）テキストの安全な匿名化に関する研究

研究課題名（英文）Safety Text De-identification

研究代表者
荒牧　英治（Aramaki, Eiji）

奈良先端科学技術大学院大学・研究推進機構・特任准教授

研究者番号：７０４０１０７３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：電子カルテの普及とともに，大量に臨床データが蓄積されつつあるが，未だそのデータの持つ情報を余すところなく利用した有望な研究や医療サービスは登場していない．その理由の1つに，研究対象のデータであるカルテが機微性の高い個人情報を含むものであり，流通や共有が困難なことが挙げられる．本研究では，自然言語処理技術を用いた新しい匿名化手法を活用し，カルテのテキスト情報を完全に匿名化することを狙いとしている．

研究成果の概要（英文）：The exponential growth in the amount of text data requires a method of a text de-identification. So far, most of de-identification methods detect the named entities, such as a person name, location name, IDs and so on, from the texts, and remove them. However, the named entity based methods suffer from a case that non-named entity conveys the personal information. To deal with this problem, this study proposes a new de-identification method that removes document specific expressions. By using this method, possible expression in a document appears in at least other k-1 documents.

研究分野：　自然言語処理

キーワード：　医療情報学　自然言語処理

１．研究開始当初の背景

現在，急速に医療の IT 化が進み，その結果，かつてない大量の臨床データが電子化された状態でストックされつつある．しかし，実際に大規模にカルテを共有し解析するためには，電子カルテに自然言語で入力される箇所について，これを匿名化する必要がある．このため，米国では 1996 年に HIPAA（Health Insurance Portability and Accountability Act）により，匿名化されるべき情報が明確に定義された．同時に，自然言語処理技術を用いて，De-identification と呼ばれる個人情報の自動匿名化に関する研究も行われている．

我々も 2006 年から匿名化の研究を開始し，2007 年では F β=1 値 0.98 と世界 4 位(当時)の精度で，個人情報を除去に成功している[文献 1]．この精度は，すでに人間の抽出精度と同等であるが，このような高精度であってもカルテの他施設への提供は進まない．これは個人情報の削除だけでは十分な匿名化と言えない場合があるからである．

例えば，「2012 年の小児の移植ドナー」など年間に数例しか行われない治療であった場合，個人情報の範囲外である単なる年代と術名の組み合わせから個人を特定可能な場合がある．また，筆記上の特徴から医師が判明可能である可能性もある．

[文献 1] Aramaki E, et al.: Automatic De-identification by using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC 2006.

２．研究の目的

本研究では，文章固有の表現を削除することで，文章を特定不可能とする新しいタイプの匿名化を提案する．この匿名化により，任意の文字列が最低 k 回以上出現するようにテキストの一部が削除される．筆者らはこれを＜テキストの k 匿名化＞とよび，そのアルゴリズムについて研究開発を行う．

３．研究の方法

方法の高速化（情報工学的研究；phase-1），圧縮率の数学的保証（数理的研究; phase-2），実証実験（臨床的研究; phase-3）という 3 つの観点から研究を行った．もっともネックになるのは，実際に個人情報を扱う phase3 であり，ダミーのカルテデータを用いて分析を行った[Imachi2013]．研究の結果，提案する方法は以下である．

まず，本研究で扱うテキストの k 匿名性を定義する．「テキストの k 匿名性」とは，ある文章に含まれるあらゆる表現が，他の文章にも k-1 回以上出現する状態とする.これが実現されると，どのような文字列で文章を検索しても，ヒットする件数は k 件以上となるかもしくはヒットしない状態となる．すなわち，文章特有の表現がない状態といえる．これを実現するためには， k 文をサンプルし，それらの差分をなくすという処理を，すべての文に対して繰り返し行えばよい．しかし，この方法を素朴に行うと，多くの文字が削除されてしまう．よって，削除する文字長を少なくするために，サンプルした k 文同士の共通文字列を多くすればよい．

提案手法（k=2 の場合）は以下の手続きを踏む．コーパス全体が n 文あるとする．それぞれの文（S_x と表記する）について，以下の処理を行う．まず，対象となる文（S_x）ともっとも類似した文を探す．類似度には様々な尺度が考えられるが，本研究では非共通文字列 (Uncommon Subsequence) の長さの逆数を用い，この結果，得られた最も類似した文を S_y とする：

$$S_y = \mathrm{argmax}_{i=1..n}\ \mathrm{sim}\ (S_x, S_i)$$

次に，S_x と S_y の差分（非共通文字列）を S_x から削除する．例えば，以下の 2 文が与えられた際，以下のような■部分が削除される．

S_x	私は大学に行った	■は大学に行った
S_y	彼は東京大学に行った	（未処理）

これで，S_x に対する処理は終わりとなる．以降の処理で， S_y に対しても S_x が類似した文となった場合は，以下のように差分が削除される．

S_x	私は大学に行った	■は大学に行った
S_y	彼は東京大学に行った	■は■■大学に行った

また，新たな文である S_z「大学に遊びに行った」と最も類似した文が S_x となった場合，S_x はそのままに，S_z のみを修正する：

S_z	大学に遊びに行った	大学に■■■行った

以上の処理により，以下の３文が得られる：

S_x	■は大学に行った
S_y	■は■■大学に行った
S_z	大学に■■■行った

■の部分は完全に削除してもよいし，何かの文字列が入っていたことを示すものとして，「*」などの代替文字列と置換してもよい．これらの文は k=2 匿名性を達成しており，以下のように，どのような部分文字列も 2 回以上マッチする（「*」はワイルドカードを示す）．

この模擬コードを以下に示す：

```
Input: set of sentences (S1..n )
Output: set of sentences (S'1..n)
for each x (1..n){
  next if (get_author (Sx) == get_author(Si));
  Sy =argmaxi=1…n (sim(Sx, Si) );
  Sx =Remove_uncommon_substring (Sx, Sy);
}

sub get_author (S){
    // Get author of sentence S
    return the author of S;
}
sub sim(Sa, Sb){
   // Calculate similarity between Sa & Sb
   return the similarity;
}
sub Remove_uncommon_substring (Sa, Sb){
   // Get differece between Sa and Sb
   Sa' = remove the diff from Sa;
   Return Sa' ;
}
```

コーパスの各文(n)について，もっとも類似した文を n 文から探索するので，素朴な計算量は $O(n^2)$となる．ただし，事前に類似した文字列のインデックスを構築しておくことで，高速化が可能となる．このための効率のよい方法は，Locality Sensitive Hashing，DivideSkip，CPMerge アルゴリズムによって実現可能である．なお，本稿の実験では高速化を実装せずに行った．

４．研究成果

削除量の保証，高速化を行い，さらに実際の医療文章に適応する実証実験を行った．まず，匿名化技術の高速化を実現については，これを実装したデモ・システム「匿名コピー」を開発した．本研究は，第 33 回医療情報学連合大会研究奨励賞を受賞するなど，高い注目を浴びた[宮部 2013]．

統計的性質については，残念ながら本研究の期間内に十分な議論を行うことができなかった．ただし，コーパス・サイズが大きくなるにつれ，保存率は上昇している性質を実験的に明らかにした(図 1)．すなわち，大きなテキストを扱えば扱うほど，匿名化されずに利用できる割合も増える．もし，これが常に成り立つのであれば，小規模のコーパスで匿名化を行い保存率を求めることで，大規模で行った場合の保存率の下限を知ることができる．

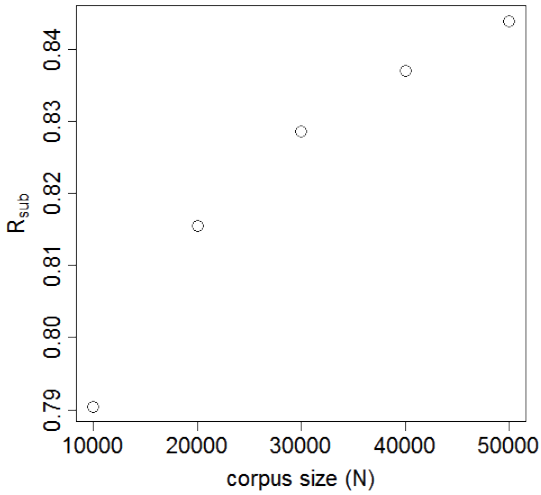

図 1: コーパスの部分保存率(R_{sub})とコーパス・サイズ（N 文）．

以上，まとめると，本研究では，新しい匿名化手法を提案した．従来の匿名化は個人名，施設名などの固有表現を削除するものであった．一方，本研究の匿名化はテキストをどのような文字列で検索しても，ヒットするテキストを k 件以上とする技術である．この匿名化は，固有名以外の部分から個人を特定できてしまうという従来の問題をクリアできる．また，匿名化された状態は明確であり，匿名化に失敗するというリスクはない．ただし，匿名化にあたって多くの文が削除されてしまうと用途に適さない可能性がある．そこで，本研究は実用化するためには必須となるどの程度の文字列が削除されるかの推定法もあわせて提案した．

今後の課題としては以下がある．

【高速化／並列化】現状のアルゴリズムでは，10 万文程度のテキストの匿名化に数日を要する．効率のよい計算法が望まれる．

【実証試験】 匿名化のような実用的な技術は速やかに運用され，運用例をつくることが重要である．運用実績が望まれる．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 3 件）

1. 荒牧英治，久保圭，仲村哲明，島本裕美子，宮部真衣: Augmented Copy: 自然言語処理を重畳するコピー機,インタラクション 2014, 未来館（東京），20140227.
2. 宮部真衣，森田瑞樹，荒牧英治: 医療テキストを対象とした言語処理実装シ

ステムとそのデータ構造，第 33 回医療情報学連合大会，神戸国際会議場（兵庫県），20131121.
3. Hiroto Imachi, Mizuki Morita, Eiji Aramaki: NTCIR10 MedNLP Baseline System, In Proceedings of NTCIR-10, NII（Tokyo），20130618.

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

〔その他〕
ホームページ
http://mednlp.jp

６．研究組織
(1)研究代表者
荒牧　英治 （Aramaki Eiji）（奈良先端科学技術大学院大学・研究推進機構・特任准教授）
研究者番号： 70401073

(2)研究分担者
森田　瑞樹（Morita Mizuki）（岡山大学・医学部附属病院・准教授）
研究者番号： 00519316