

一般口演 | セキュリティとプライバシー保護

一般口演20

情報セキュリティ・プライバシー

2021年11月21日(日) 15:10 ~ 16:10 G会場 (2号館3階232+233)

[4-G-3-04] プライバシーに配慮した診療データマスキングのための非構造化文書の分析と処理精度向上の検討

*向井 まさみ^{1,2}、江口 良太³、中川 大輔³、長谷川 裕明³、田中 勝弥^{2,1}、三原 直樹^{2,1}（1. 国立がん研究センター中央病院医療情報部, 2. 国立がん研究センター情報統括センター, 3. ファインデックス株式会社）

*Masami MUKAI^{1,2}, Ryota Eguchi³, Daisuke Nakagawa³, Hiroaki Hasegawa³, Katsuya Tanaka^{2,1}, Naoki Mihara^{2,1}（1. 国立がん研究センター中央病院医療情報部, 2. 国立がん研究センター情報統括センター, 3. ファインデックス株式会社）

キーワード：NLP, Morphologic analysis, Data Processing, Masking

【背景・目的】患者の症状や予後に関連する情報が含まれている非構造化形式の診療データ解析する試みが多数実施されている。解析には実際のデータを確認する必要があるが、プライバシーの配慮の観点から診療データの加工を行う必要がある。我々は、加工ルールを利用したフリーテキストのマスキング処理について検討している。今回、対象となる診療データの傾向を整理したので報告する。【方法】直近5年間で当院を受診した患者の診療データのうちフリーテキスト形式の診療情報として、プログレスノート、手術レポート、病理レポート等6種類の記録を対象とした。これらの記録に対して形態素解析による固有名詞（人名、地名）を抽出し全記録に対する割合を確認した。【結果】対象症例約3万例について約630万件の記録が抽出された。固有名詞の割合は手術レポート(0.55%)、プログレスノート(0.51%)の順に高かった。プログレスノートの固有名詞含有率は、医師(0.36%)、看護師(0.13%)の順に高く、中央部門スタッフの記載はほぼ0%であった。全記録のなかで、形態素解析の要素が固有名詞であるものをマスキングした記録は39%、記録文書の特徴を捉えた加工ルールに基づく加工を行ったのは18%であった。【考察】患者に直接アクセスしながら記載するものや担当者名を明記する記録に固有名詞出現率が高率なのは診療フローを考えると矛盾がない。形態素解析の要素判定だけでは十分なマスキングを行うことはできず、文書ごとに個別にルールを生成し適用することで精度があがった。しかし、マスキングが必要な用語と薬品名等に含まれるマスキング不要の用語を正しく区別する必要があるため、そのためには辞書の充実が求められることがわかった。【結語】適切な辞書の生成がマスキング処理の精度向上に貢献することが示唆された。今後、辞書充実と加工ルールへの適用方法を検討したい。

プライバシーに配慮した診療データマスキングのための 非構造化文書の分析と処理精度向上の検討

向井まさみ^{*1*2}, 江口 良太^{*3}, 中川 大輔^{*3}, 田中 勝弥^{*2*1}, 長谷川 裕明^{*3}, 三原 直樹^{*2*1}

*1 国立がん研究センター中央病院医療情報部、*2 国立がん研究センター情報統括センター、
*3 ファインデックス株式会社

Proposal for privacy-conscious masking of free text information in electronic medical records

MUKAI Masami^{*1*2}, EGUCHI Ryota^{*3}, NAKAGAWA Daisuke^{*3}, TANAKA Katsuya^{*2*1},
HASEGAWA Hiroyuki^{*3}, MIHARA Naoki^{*2*1}

*1 National Cancer Center, Hospital, Division of Medical Informatics, Tokyo, JAPAN,

*2 National Cancer Center, IT Integration and Support Center, *3 FINDEX Inc., Tokyo, JAPAN

There have been many attempts to analyze unstructured medical data that contain information related to patients' symptoms and prognosis. Although it is necessary to check the actual data for analysis, it is necessary to process the medical data from the viewpoint of privacy consideration. We have been studying the masking process of free text using processing rules. In this article, we report on the trend of the medical data to be processed. In this paper, we report on the trends of the medical data of the patients who visited our hospital in the last five years, including six types of free text records such as progress notes, surgical reports, and pathology reports. We extracted proper nouns (names of people and places) from these records by morphological analysis and confirmed their ratio to all records. Results] About 6.3 million records were extracted from about 30,000 cases. The percentage of proper nouns was higher in surgical reports (0.55%) and progress notes (0.51%), in that order. The percentage of proper noun content in progress notes was higher in the order of physicians (0.36%), nurses (0.13%), and almost 0% of central department staff. Among all records, 39% of the records were masked when the morphological analysis element was a proper noun, and 18% were processed based on processing rules that captured the characteristics of the recorded documents. The high rate of occurrence of proper nouns in records that are written with direct access to the patient or that specify the name of the person in charge is consistent with the flow of medical treatment. Morphological analysis alone cannot provide sufficient masking, and the accuracy was improved by generating and applying individual rules for each document. However, it is necessary to correctly distinguish between terms that require masking and terms that do not require masking, such as those contained in drug names, etc. For this purpose, it was found that an enhanced dictionary is required. The results suggest that the generation of an appropriate dictionary contributes to the improvement of the accuracy of the masking process. In the future, we would like to study how to enrich the dictionary and apply it to processing rules.

Keywords: NLP, Morphologic analysis, Privacy, Data Processing, Masking

1. 背景・目的

病院における診療情報は、一義的には患者の診療記録を正確に記載することが目的とされているため、いわゆる臨床研究への二次利用を想定した構造化された形式での入力・コードの利用等のルールが医療機関内で確立・周知徹底されているとはいえない状態にあると考えられる。一方、リアルワールドエビデンスの創出と活用には、様々な医療情報を統合解析する際の礎となるデータの信頼性の担保は不可欠である。米国をはじめ欧米諸国においては、リアルワールドデータの利活用の積極的な推進を図っているところであり、わが国においてもリアルワールドデータの利活用の本格化が急務となっている。そこで、本邦においても、AMED(国立研究開発法人日本医療研究開発機構)の「医療技術実用化総合促進事業」による、病院情報システム内の医療情報データの品質確保メカニズムを病院内に構築するとともに、あわせて臨床研究中核病院間でのデータ相互利用を可能にする標準化を図ることを目的としたプロジェクト「臨中ネット」が開始されるなど、今後複数の施設、研究者によるデータ利活用が進められようとしているところである。

しかしながら、解析対象となる診療データには、患者個人の機微な情報も多く含まれており、「人を対象とする研究の倫理指針」に基づいての研究であること、対象となる患者の研究参加の拒否機会を保証していることから、その前提として、解析目的を損なわない範囲でのデータの適切な匿名加工処理が重要な要件となっている。現状ではそのデータ加工処理は、個々の研究者のリテラシーや技術力により実現されており、データの品質が一定ではないこと、匿名化されたデータの保証に対して客観的な視点での検証方法が存在していないことが課題となっている。

我々は、病院情報システム(電子カルテ、医用画像管理システム、医事会計、部門システム群)に格納されている診療情報に対して匿名加工処理を施すアルゴリズムの開発を行うこととした。本研究のアウトラインを図1に示す。

先行の研究では、診療データの自由記載文を対象に実行した結果、記載が統一されている文章で、かつ、正しい用語の記載であれば、加工ルールを用いることでマスキング処理が可能であることと、適切な前処理を行うことで処理精度の向上が期待できることが示唆された。[9]

今回マスキング処理にかける前段階の処理の可否に対して検討を行うために、そもそも診療データのうち自由記載文にどのような用語が用いられており、マスキング処理がどのようにかかっているかを整理したので報告する。なお、本研究は、国立がん研究センター研究倫理審査委員会により承認されている(研究課題番号:2020-247)。

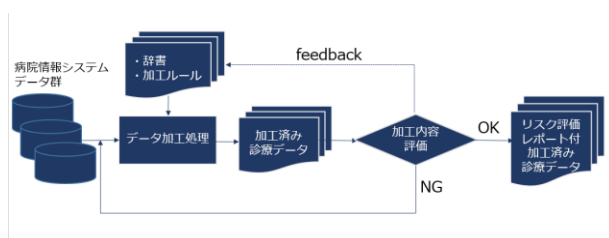


図1 匿名加工処理アルゴリズム開発のアウトライン

2. 方法

直近5年間で当院を受診・治療した患者の診療データのうち、フリーテキストデータをもつ診療情報を対象とした。対象とする診療データは、電子カルテの記事(医師の SOAP、看護記録等)、服薬指導、栄養指導、放射線診断レポート、病理レポート、手術記録の6種類とした。これらの記録に対して形態素解析による固有名詞(人名、地名)を抽出し全記録に対する割合を確認した。形態素解析にはフリーのライブラリである“GiNZA”を使用した。[1][2][3][4][5][6][7][8]

次に、マスキングの手法として先行研究で試行した①形態素解析による固有名詞を選択的に処理、②情報システムが持つ患者 ID および患者氏名による処理、③固定のキーワードを用いた正規表現による処理、の3つの処理のうちいずれにより処理されるのかを確認した。

3. 結果

対象症例約3万例について約630万件の記録が抽出された。固有名詞の割合はプログレスノート(2.1%)、手術レポート(1.8%)、の順に高く、もっとも低かったのは放射線レポート(0.9%)であった(表1)。全記録のマスキング処理を施したなかで、形態素解析の要素が固有名詞であるものをマスキングした記録は66%、患者情報を元にした加工は3%、記録文書の特徴を捉えた加工ルールに基づく加工を行ったのは30%であった。

5. 考察

医師の記事記載のように患者に直接アクセスしながら記載する記録や、フォーマットの中で術者や各担当者名を明記す

る手術記録のように病院職員の名前を明記する記録に固有名詞出現率が高率なのは診療フローを考えると矛盾がない。

加工箇所の割合に差があることから、データ種別により、マスキング対象の単語の用い方には差があることがわかった。形態素解析を文書にかけるのは処理時間がかかるので、文書種別と求められるマスキングの精度によっては処理パターンによるマスキングだけを行うことで効率的に処理を行うことができる可能性があると考えられる。

先行研究の結果とあわせて考えると、薬品名や病名に姓名にあたる文字列が使われている場合には、辞書がない状態でマスキング処理を行うと形態素解析の要素判定だけでは過剰なマスキング処理が行われてしまう。実際にマスキング処理を適用する場合にはマスキング対象外をまとめた辞書を導入する機能は必要と考えられる。

6. 結語

マスキング処理を行うべき単語を観点に、診療データの中のフリーテキストを記録種別ごとに確認した結果、内包する用語には差があることがわかった。このことから、求められるデータ加工の精度により、適用すべきマスキング処理を選択することを視野にいれることが処理スピード向上に貢献することが示唆される。

今後、先行研究でわかっている辞書の充実およびマスキング処理の加工ルールへの適用方法を検討したい。

参考文献

- 1) 形態素解析結果ライブラリ”GiNZA”, megagonlabs.
[<https://megagonlabs.github.io/ginza/>]
- 2) ICD10 対応標準病名マスター, MEDIS-DC, Ver.5.04
[<https://www2.medis.or.jp/stdcd/byomei/download2019.html>]
- 3) 臨床検査マスター, MEDIS-DC, Ver.2.73,
[<https://www2.medis.or.jp/master/kensa/index.html>]
- 4) 医薬品 HOT コードマスター.MEDIS-DC,2020年7月31版
[http://www2.medis.or.jp/mousikomi/an200607.cgi?hot_h20200731.zip]
- 5) 一簡易版一漢方辞書, 東亜医学協会,
[<http://square.umin.ac.jp/aeam/siryouko/kanibankampodic.htm>]
- 6) DMiME 医学用語変換辞書, OSDN, Ver.-1.1.1,
[<https://ja.osdn.net/projects/dmime/>]
- 7) 人を対象とする医学系研究に関する倫理指針ガイダンス, 厚生労働省, [<https://www.mhlw.go.jp/file/06-Seisakujouhou-10600000-Daijinkanboukouseikagakuka/0000166072.pdf>]
- 8) 田木真和. 個人情報の定義. 医療情報第6版「医療情報システム編」. 篠原出版新社, 2020 : 381-386
- 9) 向井まさみ, 田中勝弥, 三原直樹, プライバシーに配慮した電子カルテのフリーテキスト情報のマスキング処理についての検討. 第40回医療情報学連合大会

表1. データ種別ごとの形態素解析の結果

	服薬指導		栄養指導		手術レポート		病理レポート		プログレスノート		放射線レポート	
データレコード数	177,954		23,336		17,583		172,405		5,637,157		332,777	
感動詞	4,639	0.007%	2,436	0.030%	1,654	0.016%	645	0.004%	152,783	0.022%	3,195	0.008%
記号	9,692	0.015%	476	0.006%	1,536	0.015%	4,187	0.025%	166,254	0.024%	3,924	0.010%
形状詞	537,413	0.831%	164,344	2.055%	66,420	0.651%	130,038	0.776%	6,803,858	0.968%	298,774	0.737%
助詞	5,528,406	8.551%	1,586,610	19.836%	1,668,577	16.364%	2,465,688	14.715%	67,385,992	9.590%	6,387,382	15.760%
助動詞	1,060,248	1.640%	274,151	3.427%	382,531	3.751%	906,041	5.407%	17,558,390	2.499%	3,659,595	9.029%
接続詞	37,269	0.058%	8,569	0.107%	18,618	0.183%	49,656	0.296%	531,826	0.076%	96,501	0.238%
接頭辞	148,788	0.230%	32,096	0.401%	65,149	0.639%	124,490	0.743%	2,587,850	0.368%	201,659	0.498%
接尾辞	1,228,496	1.900%	154,147	1.927%	175,415	1.720%	576,890	3.443%	12,424,423	1.768%	661,468	1.632%
代名詞	32,184	0.050%	8,870	0.111%	11,070	0.109%	13,175	0.079%	597,081	0.085%	23,219	0.057%
動詞	2,474,537	3.827%	764,025	9.552%	715,927	7.021%	1,234,191	7.365%	31,249,707	4.447%	2,896,011	7.145%
副詞	289,193	0.447%	86,210	1.078%	29,614	0.290%	26,950	0.161%	2,836,049	0.404%	104,808	0.259%
補助記号	17,347,145	26.830%	1,393,907	17.427%	1,776,338	17.421%	3,256,532	19.434%	161,680,670	23.010%	6,231,274	15.374%
名詞-固有名詞	1,007,303	1.558%	75,348	0.942%	185,351	1.818%	216,078	1.290%	14,436,870	2.055%	348,649	0.860%
名詞-固有名詞以外	34,905,160	53.987%	3,438,814	42.993%	5,077,767	49.798%	7,737,847	46.178%	383,404,140	54.566%	19,516,291	48.152%
連体詞	44,152	0.068%	8,632	0.108%	20,800	0.204%	14,238	0.085%	831,741	0.118%	97,604	0.241%
合計	64,654,625		7,998,635		10,196,767		16,756,646		702,647,634		40,530,354	

表2. 記録種別ごとの匿名化処理箇所と適用された処理方法

	記録件数	匿名化 レコード数	形態素数	匿名化 加工箇所	匿名化種別					
					Morphological		PatientInfo		Regex	
服薬指導	177,954	64,874	64,654,625	74,441	22,603	30.4%	7	0.0%	51,831	69.6%
栄養指導	23,336	16,556	7,998,635	32,507	3,834	11.8%	876	2.7%	27,797	85.5%
手術レポート	17,583	11,075	10,196,767	31,477	20,924	66.5%	265	0.8%	10,288	32.7%
病理レポート	172,405	34,600	16,756,646	60,008	14,030	23.4%	2,456	4.1%	43,522	72.5%
プログレスノート	5,637,157	1,498,669	702,647,634	3,470,599	2,430,769	70.0%	110,670	3.2%	929,160	26.8%
放射線レポート	332,777	53,482	40,530,354	105,420	14,254	13.5%	0	0.0%	91,166	86.5%
合計	6,361,212	1,679,256	842,784,661	3,774,452	2,506,414	66.4%	114,274	3.0%	1,153,764	30.6%

*1) Morphological : 形態素解析より、名詞-固有名詞-人名、名詞-固有名詞-地名と判断されたものをマスキング
PatientInfo : 電子カルテ登録上の患者番号、氏名等のテキストデータでマッチング・マスキング
Regex : 正規表現でマッチング・マスキング