厚生労働科学研究費補助金（臨床研究等 ICT 基盤構築研究事業）
分担研究報告書

退院サマリの自動生成に向けた電子カルテの自動分析

研究分担者　狩野 芳伸
（静岡大学 情報学部 行動情報学科 准教授）

研究要旨

入院患者の退院に際し、医師は入院中に記載したカルテ等の情報から退院サマリを作成する必要がある。この退院サマリを自動的に生成することが出来れば、臨床現場の負担を下げることが出来ると共に、医療の質に貢献することが期待される。そこで、本研究分担では、退院サマリの自動生成に向けたテキストの分析についての研究を行った。

そこで、本研究分担では、退院サマリの自動生成に向けた研究アプローチの検討に取り組んだ。まず、文献調査と医師へのヒアリングに基づき、良質な退院サマリに求められる要件について定性的な検討を行った。同時に、実際の退院サマリを対象とした分析を行い、要約過程に関する知見を整理した。さらに、一般的な文書の要約手法と入院カルテの要約手法について文献調査を行った。

その結果、退院サマリの分析枠組みと退院サマリの生成モデルを兼ねた CASE モデルと証するモデルを構築することが出来た。その上で、今後、本モデルが示唆する特性の異なる 4 つの要約処理が出力した候補文集合を退院サマリの下書きとして提示し作成支援するツールのプロトタイピングが望まれることを示した。

カルテの処理にあたっては、事前に匿名化が必要となる。匿名化作業を自動化するための匿名化ツールの実装と性能向上に取り組んだ。そのために、既存の正解付き模擬カルテデータに加え、別のダミーカルテデータセットに対し匿名化のためのアノテーション付与を行い、これらを用いてルールベースおよび機械学習による匿名化ツールの実装と性能検証を行った。

サマリ生成にあたっては、対象とするカルテやサマリのドメイン、すなわち診療科や疾患により、サマリ生成に必要となる情報が異なると考えられる。サマリと対応する電子カルテの履歴データについてクラスタリングを行い、どのようなタイプのサマリやカルテがどう類似しうるかの分析を行った。

１．はじめに

本研究では退院サマリの自動生成を目指している。すなわち、入院中の記録である電子カルテを中心とする患者の履歴を入力とし、その患者の退院時の「まとめ」にあたる退院サマリを出力とするシステムの構築である。

まず、電子カルテという個人情報を扱うことから、厳重なセキュリティ環境が必要である一方、現実的に研究が遂行可能な環境の構築が必要である。

電子カルテの処理にあたっては、自然言語処理によるテキスト処理が必須である。ひとつには、個人情報保護の観点から匿名化処理が必要となる。そのうえで、電子カ

ルテデータにおける個別日時の情報（以下、履歴と呼ぶ）とサマリの間にカルテの種類に応じてどのような関係がありうるか、分析を行い、サマリの自動生成を試みた。

２．セキュアかつ効率的な研究環境の整備と運用

本研究の遂行にあたり、セキュアかつ効率的な研究環境の構築を行った。実行環境を仮想マシンとし、実行環境そのものを遠隔送信し、現地で容易に実行できるようにした。ただし、現地では秘匿すべきデータは別サーバに格納し、ソフトウェアからは一時的なアクセスとして情報を残さないようにすることでセキュリティを確保した。本年度は、このシステムを運用して研究を行った。

３．模擬カルテとアノテーションに基づく退院サマリ考察

他の分担研究により、本年度模擬カルテの提供と、その模擬カルテに基づいた、退院サマリ作成を考慮したアノテーション付与が行われた。

サマリを要約ととらえると、一般的な自動要約技術が適用できそうにも思われる。多くの自動要約は、トピックの統計的な解析を行ったうえで、文書中で重要なトピックが含まれるものを残す、という手法が骨格になっている。しかし退院サマリでは、統計的に重要でない、文書集合中で共通して頻出するトピックであっても、サマリとして残すべきことが多々ある。

また、入力にあたる電子カルテの文章中にない文章や表現が、サマリにどのくらい含まれているかという問題がある。入力のサブセットでよいのであれば、切り貼りの範囲内におさまるが、現実には言い換えに始まり内容的にも新規な文章の挿入がありうる。

分担研究のデータによると、入院までの経過については７割以上のサマリ文がカルテの文章ほぼそのままであった。このことは、医師がサマリを作成する際に文の複製を使用しており、分量的な減少もあまりみられないことから、内容的にもあまり変更を必要としていないことを示唆している。ただし、入院前の記述は他の医師からのお願いの形式をとっており、そのままでは主語や言葉遣い、時制などが不適切なので、そうした部分の変換が必要かもしれない。

入院中の経過については、２～３割程度がそのままの文であった。入院中のカルテの記載は文を完成させず断片的なスタイルのことが多く、一方サマリではきちんとした文にするため多かれ少なかれ文生成の要素が必要と思われる。また、医師本人の記録なので、振り返ることで要素を追加したり、整理したりすることが想像される。

４．カルテとサマリの間の類似性に関する予備的実験

どの文を（一部であっても）サマリとして取り入れるかを自動判定する際、最も基礎的な要素は単語の共通性になる。そこで、予備的実験として、共通する単語の分布を測定した。対象を内容語のみにすると、当然ながら「入院」「退院」「病名」など、入院カルテにおける一般的な単語が上位にみられた。一方で、部位、症状、病名、薬品名、単位などを表す単語も上位にきており、最初の手掛かりになろうと思われる。

５．自動的な匿名化

電子カルテの実データに対し処理を行うには、まず匿名化が必要となる。具体的には、個人氏名、年齢、住所、日付、医院名

など個人を特定しうる情報の自動抽出である。

匿名化の研究は長らく行われているものの、特に日本語医療分野の匿名化は利用可能なリソースが少ないこともあり、研究の数は限られている。利用可能なリソースとしては、NTCIR MedNLP 匿名化タスクで配布された模擬カルテコーパスとそこに付与されたアノテーションが挙げられる。このとき評価に使われた正解アノテーションは期間限定で利用不能となっており、残念ながら直接的な比較を行うことはできない。

そこで、研究班内で作成されたダミーカルテを対象に、新たに匿名化のための正解アノテーションを付与し、学習及び評価に用いた。付与するアノテーションは基本的に MedNLP タスクにおけるものと同種とした。すなわち、年齢・個人名・医院名・性別・時間表現である。

これらのデータを用いて、自動匿名化ツールのプロトタイプ実装を行い、性能を検証した。手法としては、ルールベースのものと機械学習によるもの、およびそれらの混合を試みた[Kajiyama 18]。前述の理由から、MedNLP タスクにおける先行研究との直接比較はできないが、概ね同等程度の性能が得られたと考えられる。また手法としては、文字ベースの LSTM（Long-Short Term Memory）を用いたものが最も安定して高性能であった。ただし、いずれのデータセットも模擬カルテの作成コストが大きく、サンプル数が不足している。そのため、end-to-end の機械学習のみでは性能が不十分であり、ルールベースの併用が必要である。

６．カルテとサマリの間の類似性に関する実験

サマリを要約ととらえると、一般的な自動要約技術が適用できそうにも思われる。多くの自動要約は、トピックの統計的な解析を行ったうえで、文書中で重要なトピックが含まれるものを残す、という手法が骨格になっている。しかし退院サマリでは、統計的に重要でない、文書集合中で共通して頻出するトピックであっても、サマリとして残すべきことが多々ある。

また、入力にあたる電子カルテの文章中にない文章や表現が、サマリにどのくらい含まれているかという問題がある。入力のサブセットでよいのであれば、切り貼りの範囲内におさまるが、現実には言い換えに始まり内容的にも新規な文章の挿入がありうる。

研究班内のデータによると、入院までの経過については 7 割以上のサマリ文がカルテの文章ほぼそのままであった。このことは、医師がサマリを作成する際に文の複製を使用しており、分量的な減少もあまりみられないことから、内容的にもあまり変更を必要としていないことを示唆している。ただし、入院前の記述は他の医師からのお願いの形式をとっており、そのままでは主語や言葉遣い、時制などが不適切なので、そうした部分の変換が必要かもしれない。

入院中の経過については、２～３割程度がそのままの文であった。入院中のカルテの記載は文を完成させず断片的なスタイルのことが多く、一方サマリではきちんとした文にするため多かれ少なかれ文生成の要素が必要と思われる。また、医師本人の記録なので、振り返ることで要素を追加したり、整理したりすることが想像される。

このようにさまざまな要素があるが、現実には診療科や疾患のタイプによって、類似性の高いものとそうでないものがありうる。そこで、国立病院機構内の実際のカルテを対象にクラスタリングを行い、カルテに記載された ICD や DPC といったラベルを含めて類似性の解析を行った。

７．サマリの自動生成と評価

前節までの結果を踏まえ、退院サマリの自動生成とその評価を行った（論文投稿予定）。退院サマリの自動生成にあたっては、extractive な処理を行うこととし、電子カルテ内の各文についてサマリに含めるべきか否かの判断を行った。判定には、文末表現に着目する手法、文ベクトルを生成して文間の類似度で決定する方法などいくつかの手法を行い、文一致率、単語一致率、ROUGE など異なる指標での評価を行った。結果、いずれの手法でもある程度の一致率を達成しうることがわかった。今後、より実用的な性能の達成のためには、表記揺れの吸収、さらに大規模なデータの利用による学習性能の向上などの研究が考えられる。

[Kajiyama 18] Kajiyama, K. Horiguchi, H. Okumura, T. Morita, M. Kano, Y., 2018. De-identifying Free Text of Japanese Dummy Electronic Health Records. The Ninth International Workshop on Health Text Mining and Information Analysis(LOUHI2018) (p. 65).