Vol.2022-CLE-38 No.12
2022/11/5

既存手法組み合わせによる 網羅的な診療録監査業務の課題解決について

行田優衣 [1]

概要：診療録各種監査督促リスト作成については，一部のタスクを自動化することが可能となった．しかし，退院時要約 2 週間以内完成率については，医師は多忙なため毎月変動があり，診療録管理体制加算（Ⅰ）の算定要件に満たない月も見受けられた．また，質的監査については，外部監査機関の対象項目となっており，診療録の算定要件にも該当するものもある．厚生労働省 Jobtag では，英語や医学用語のまじった診療録の内容を正しく理解する能力の必要性についても報告されている．今回は，現実に存在するタスクの課題解決のために，3 つのテーマに関して既存手法の組み合わせについて考えた．診療録（退院時要約）の文書生成， 網羅的な質的監査，英語・医学用語翻訳に関するものである．個人情報保護法の観点から医療機関のデータは使用していないが，上記の３つの課題に関しては，既存手法をうまく組み合わせることによって課題解決に繋がると考える．

キーワード：文書生成，網羅的な質的監査，英語・医学用語翻訳[**]

Based on a combination of previous research methods Solving the Problem of Comprehensive Medical Record Auditing Work

YUI YUKUTA[†1]

Abstract: For preparation of various audit dunning lists of medical records, It has become possible to automate some tasks, but the completion rate within 2 weeks at the time of discharge varies from month to month due to the busy schedule of doctors, and there were months when the calculation requirements for the medical record management system addition (I.) were not met. In addition, qualitative audits are subject to external auditing organizations, and some of them also fall under the calculation requirements of medical records. The Ministry of Health, Labour and Welfare's Jobtag also reported the necessity of the ability to correctly understand the contents of medical records mixed with English and medical terms. In order to solve the problems of tasks that exist in reality, we considered a combination of previous research methods on three themes: Document generation of medical records (summary at the time of discharge), comprehensive qualitative audit, This is about the translation of English and medical terms. This time, we do not use data from medical institutions from the viewpoint of the Act on the Protection of Personal Information, but we believe that the above three problems will lead to problem solving by skillfully combining previous research methods. [**]

Keywords: Document generation, comprehensive qualitative audits, English and medical term translation [**]

1. はじめに

保険診療は，健康保険法等の各法に基づく，保険者と保険医療機関との間の「公法上の契約」に基づいている．保険医療機関は，健康保険法等で規定されている保険診療のルール（契約内容）に従って，療養の給付及び費用の請求を行い，保険医は保険診療のルール（契約の内容）に従って，療養の給付及び費用の請求を行っている．ルールに従って行われていない場合は，行政指導等の対象となる．保険診療として診療報酬が支払われる条件の一つに,「診療報酬点数表に定められたとおりに請求を行っていること」があげられる．そして，診療報酬請求の根拠は「診療録」にあるとされている[1].

病院情報システムには，病院の基幹業務を遂行するために一次データを保管する業務用データベース（以下 DB）が作成される（例：患者基本情報 DB，外来診療費 DB，入院診療費 DB）．それらのデータを活用して，分析を目的とする病院組織を横断的に集約した DB の集合が DWH である．DWH は基幹業務系システムとは別系統で構築されるので，データの保存の信頼性の向上する．DWH は大量の時系列データを蓄積しており，情報の宝の山である．膨大な情報を蓄積している DWH から新しい知見や発想を導き出すデータ処理の過程をデータマイニングと呼ぶ．さらに，この大量データから法則性の検証を行うために仮説を立て正当

1 独立行政法人国立病院機構　福山医療センター
Medical information management officer
NATIONAL HOSPITAL ORGANIZATION
FUKUYAMA MEDICAL CENTER

性を調べるのがオンライン分析である[2].

経営分析に必要なデータをこの DWH に時系列で蓄積しておいて，適宜，経営管理指標 DB として活用することが可能になった．ほかにも病院運営の意思決定，経営判断および将来計画の作成など広範囲にわたる強力な支援機能を発揮する．さらに，医療安全対策，部門別原価計算，在庫管理，人事管理など多様な病院業務の情報処理，情報解析に有効である[2].

しかし，病院経営上必要なデータ抽出，医学研究のための診療情報，医療訴訟に耐えうる診療録作成のための各種監査データは，その都度レセプトデータ（EF ファイル，D ファイル等），DWH から抽出する必要がある．抽出後，さらに様々な依頼条件に従ってデータを加工する必要がある．まず，病院経営上必要なデータ抽出について，診療報酬請求上必要なデータは「実施オーダ」が自動的に発行されることもある．しかし，当院のように麻酔管理料（Ⅰ）等の医学管理料はテンプレートを手動で医師事務作業補助者が作成している病院も多いと思われる．そのテンプレートを算定担当者が確認後，医学管理料の算定が行われるため，カルテ記載・算定漏れがある場合は，診療情報管理士が電話で担当者へ依頼している．患者数が多く，手術当日退院の患者に関しては，カルテ記載，算定が退院までに間に合わないこともあり問題となっている．

近年，医学管理料算定フォローシステムの在り方について，報告されている．システム化するメリットとして，矯正する力（漏れ防止-気づき）と平準化する力がある[3].

「電子カルテにおけるインフォームドコンセント自動監査システム」の構築に記載されている今後の課題解決として,「診療情報管理士の主な指摘事項を類型化し問題のある記載の指摘パターンを推定する」ことがあげられている[4]．診療情報管理士等のデータ作成者，監査担当者が多忙である場合は，依頼者からの締め切りに追われ，時間が掛かる等の問題が発生する．そこで今回は，現実に存在するタスクの課題解決のために，3 つのテーマに関して既存手法の組み合わせについて考えた．診療録（退院時要約）の文書生成， 網羅的な質的監査，英語・医学用語翻訳に関するものである．個人情報保護法の観点から医療機関のデータは使用していないが，上記の３つの課題に関しては，先行研究手法をうまく組み合わせることによって課題解決に繋がると考える．

このようなシステムを構築することにより,医学研究のための診療情報データ作成等，他の様々な診療情報データ抽出に対応可能であると考える．

2. 背景

(1) 診療録の文書要約・分類（退院時要約等）

診療録各種監査督促リスト作成については，一部のタスクを自動化することが可能となった．しかし，退院時要約 2 週間以内完成率については，医師は多忙なため毎月変動があり，診療録管理体制加算（Ⅰ）の算定要件に満たない月も見受けられた．現在の病院情報システムでは，化学療法，放射線治療等一連の治療で繰り返し入院の場合に有効と考えられる項目は，退院時サマリ登録画面にある「前回記載複写ボタン」を手動でボタンを押すと表示される．また，入院までの経過・入院時現症・既往歴・アレルギー・臨床経過・治療方針については，医師が指定のテンプレートを利用した場合，既存システムでは電子カルテのテンプレート機能から自動流し込み可能である．しかし，テンプレートを使用しない医師が多いことがあり，登録漏れがある場合，監査担当者が手術情報や退院処方等は，それぞれの追加ボタンを押し，表示されたオーダを手動で貼付しなければならない.また,日付誤り等の誤字が見受けられる．日々患者カルテを一つずつ手動で開いて，量的な部分に関しては，現在導入されている病院情報システム上の理由から診療情報管理士が確認，訂正し更新しているのが現状である．

問題の解決方法として，「退院時要約作成支援ツール（DPC の EF ファイルや看護ケア実績情報を元にした診療実績表)」が開発されている[4]．このようなツールを活用して，文書生成を行うことが必要であると考える．

近年，自動問診から診断・治療支援までを視野に入れた AI システムの開発が進んでいる．それに伴いテキスト解析や紹介分の自動作成機能が開発されている．将来的には，これらの技術を応用し，診療録全体の自動作成や次に述べる過去の診療録に関する各種監査，英語・医学用語の自動翻訳に対応するシステムを構築することが望ましいと考える．上記のタスクを自動化することにより得られるデータは記載漏れや誤字脱字が減少し，将来の各種監査システムは，新たな症例に関して医師等の医療従事者が主導で入力した際に用いることが可能であると考える．

既存手法として，文章以外のデータから文章を生成する技術の中で「言語モデル方式」がある.言語モデル方式ではテンプレートを用いずに文章生成に言語モデルを活用する.具体的にはデータと文章のペアを元に深層学習など利用してデータから直接文章を生成するモデルを作成する．正しいモデルを作成できれば，データを与えることで直接文章が得られる[5].

多次元センサーデータが与えられたときに，その時系列データの中から重要な特徴を発見し，それらの情報を統計的に要約し，表現する．

(2) 網羅的な質的監査

先程述べたような自動問診・診断・治療支援システムから診療録全体の自動作成や次に説明する過去の診療録に関する各種監査，英語・医学用語の自動翻訳に対応するシス

テムが構築されると，新たな症例等，数少ない症例のみの利用となってくると予測される．

しかし，過去に手動で作成した診療録を機械学習の分類を用いて主な指摘事項を特定し集計することは，今後上記のようなシステムが構築された際にも，医学研究・医療訴訟に耐えうる診療録の作成に向けての注意点が把握可能なメリットがあると思われる．

(3) 英語・医学用語翻訳

カルテの記載は，原則日本語で行うべきである．医師以外の医療職種や事務職，さらに患者や患者家族（カルテ開示の場合)，さらに訴訟の際の証拠として採用されることを考えれば，ニュアンスも含め誤解を生じないように記載できる言語は日本人にとって日本語しかないからである[7]．

医学・医療用語には，しばしば略語が利用される．略語を使用する場合は，その略語が一定の市民権を得ていること（学会や公式文書等で利用できる程度に普及している）が前提である．少なくとも，ある病院でしか通用しない略語は用いるべきではない．また，略語に用いる元の言語は英語に限定すべきである．現在，急性期病院で複数の診療科で一般的に使用されている、病名・病態名，検査目，手技名，治療名，薬品・薬剤名等を網羅的に集めた用語集がある[7]．

3. 関連研究

3.1 深層学習を用いた時系列データの要約と分類

診療録の文書生成（退院時要約等)，網羅的な診療録質的監査について，応用可能と考えられる手法は，深層学習を用いた時系列データの要約と分類がある[8]．多次元センサーデータが与えられたときに，その時系列データの中から重要な特徴を発見し，それらの情報を統計的に要約し表現する．さらに，要約情報に対して深層学習を適用することにより，時系列データを効率的かつ効果的に分類する．実データを用いた実験では，時系列データの中からラベル分類に関する重要パターンを抽出し，高い精度で分類できることが確認された．

3.2 ニューラル機械翻訳での目的言語側の文脈の効果的な利用

文脈を考慮したニューラル機械翻訳の精度向上のため，目的言語側の前文の参照訳と機械翻訳結果の両方を文脈情報として用いる手法が提案された[12].文脈として，原言語側または目的言語側の周辺の文が利用できるが，目的言語側の周辺の文を用いる手法は翻訳精度が下がることが報告されている．目的言語側の文脈を利用したニューラル機械翻訳では，学習時は参照訳を用い，翻訳時は機械翻訳結果を用いるため，参照訳と機械翻訳結果の特徴の異なり（ギャップ）が原因の１つと考えられる．そこで，学習時と翻訳時の目的言語側の文脈情報のギャップを緩和するために，学習時に用いる目的言語側の文脈情報を学習の進行に応じて参照訳から機械翻訳結果へ段階的に切り替えていく手法が提案された．時事通信社のニュースコーパスを用いた英日・日英，および英独・独英機械翻訳タスクの評価実験により，従来の目的言語側の文脈を利用した機械翻訳モデルと比較して翻訳精度が向上することが確認された[12].

3.3 医療用語資源の語彙拡張と診療情報抽出への応用

診療記録では，多様な構成語彙の組み合わせからなる複合語が使用されるため，単純なマッチングに基づく辞書の利用では検出できない用語が存在し，語彙資源利用の効果は限定的となる．そこで，語彙資源を有効活用した用語抽出が提案された．1 点目として，資源中の用語に対して語彙制限を行うことで，用語抽出に真に有用な語彙の獲得が行われた．2 点目として， 資源から複合語の構成語彙である修飾語を獲得し，元の語彙に加えて獲得した修飾語を活用することで，テキスト中のより多くの用語を検出する拡張マッチングが行われた．結果，単純な語彙資源の利用時と比較して適合率および再現率の向上を実現し，本手法の有効性が確認された[8]．

3.4 形態論的制約を用いた未知語の自動獲得

日本語は分かち書きされないため，複数の用例を調べなければ，適切に未知語を処理できないとされている．形態論的制約を用いた未知語の自動獲得では，テキストから逐次的に語彙を獲得し，その場で形態素辞書を自動更新する枠組みと，その具体的な実装手法が提案された[10]．語彙獲得の手がかりとしては，自立語に後接する付属語列に関する形態論的制約を用いる．複数の用例における付属語列の振る舞いを調べることにより，未知語が確実に獲得できることが示された．

3.5 自動獲得した未知語の読み・文脈情報による仮名漢字変換

自動獲得した未知語の読み・文脈情報による仮名漢字変換に関する研究が行われている．内容の類似したテキストと音声から未知語の読み・文脈情報をコーパスとして自動獲得し，仮名漢字変換の精度向上に利用する手法が提案された[11]．まず，確率的な単語分割によって未知語の候補となる単語をテキストから抽出する．次に，各未知語候補の読みを複数推定して列挙する．その後，テキストに類似した内容の音声を認識させることによって正しい読みを選択する．最後に，音声認識結果を学習コーパスとみなして仮名漢字変換のモデルを構築する．自動収集されたニュース記事とニュース音声を用いた実験では，獲得した未知語

の読み・文脈情報を仮名漢字変換のための学習用コーパスとして用いることで，精度が向上することが確認された[11].

4. 要約・分類 [6]

・4.1 深層学習を用いた時系列データの要約と分類[6]

【問題定義】

ここでは,「深層学習を用いた時系列データの要約と分類」という既存手法にしたがって必要な概念について定義を行う． $X = \{X(k)\}_{k=1}^{K}$ を K 個のセンサーによって観測された長さ t の多次元時系列データとし，各センサー k は d (k) 個の測定軸 を持つとする．本研究では，このようなラベル付き多次元時系列データの集合 $D = \{Xi\}_{i=1}^{N}$ に対する分類を行う． 本研究ではまず，単一の時系列シーケンス X を m 個のセグ メント集合 S = {s1, ・・・, sm} に分割する．ここで，si は i 番目 のセグメントの開始点，終了点を含み，(つまり，si = {ts, te})， 各セグメントは重複がないものとする．次に，発見したセグメント集合を類似セグメントのグループ(レジーム: regime) に分ける．これをローカルレジームと呼び，各レジームは統 計モデル θ で表現される．また，各サンプルから得られるレジームに加えてデータセット全体での共通パターンであるグローバルレジームを検出する．最終的に，これらの要約情報 を利用した深層学習によって各時系列データを分類するとともに，時系列パターン（レジーム）の分類問題に対する重要度 $\alpha = \{\alpha 1, \cdots, \alpha m\}$ を算出する．

［定義 1］(ローカルレジーム) 多次元時系列シーケンス X が 与えられたとき，X を表現，要約する ML = {m, r, S, Θ, α} をローカルレジームと呼び，以下で構成される．

（１） セグメントの総数 m と各セグメントの位置：
S = {s1, . . . , sm}

（２） r 個のレジームを表現するモデルのパラメータ集合： $\Theta = \{\theta 1, \ldots, \theta r\}$

（３） 各セグメントの重要度：
$\alpha = \{\alpha 1, \cdots, \alpha m\}$

［定義 2］(グローバルレジーム) D を表現する全ての要約情 報 MG = {ML 1 , . . . ,ML N , ΘG} をグローバルレジームとし， $\Theta G = \{\theta 1, \ldots, \theta g\}$ をシーケンス間で共通するパターンを示す g 個の統計モデルの集合とする． したがって，本論文で扱う問題は以下のように表される．

［問題 1］ 多次元時系列データ $X \subset D$ が与えられたとき，

（１） 要約情報 ML，および，MG を抽出する

（２） X に含まれる各時系列パターン（レジーム）から特徴量 X を生成し，X に付与されたラベル y を高精度に分類する

（３） 各セグメントの分類問題に対する重要度 α を求める

結論として，本論文の目的はデータセット全体の最適な要約 情報を抽出し,要約情報に基づき生成した特徴量によってそれぞれの時系列シーケンスを分類するとともに，分類問題における時系列パターンの重要度を算出することである．ここで非常 に重要な課題は，(a) どのように X の要約情報 ML を推定するか，(b) どのようにデータセット全体のグローバルレジーム MG を決定するか，(c) 時系列パターンの重要度をパラメータ に持つ分類モデルをどのように構築し，高精度に学習するかで ある．本論文では，多次元時系列データを要約･分類するため のアルゴリズムである DeepPlait を提案する[6].

提案モデルのネットワーク構造:

(a) 特徴抽出層: 時系列データをパターンごとに分割し特徴量を生成する．

(b) CNN 層: 分 割された各特徴量におけるセンサー軸間，センサー間の関係性を抽出する．

(c) RNN 層: CNN 層で得られた特徴の時間関係を学習する．

(d) Attention 層: RNN 層で時間関係を学習するとき，時系列パターンの重要度を加味する[6].

・4.2 システムの提案

＜診療録（退院時要約）の文書生成＞

時系列データに含まれるパターンに関する情報，すなわち，パターンの変化点や種類がすべて明らかであることは稀である．収集した膨大な時系列データをすべて人手で確認し，類似パターンをラベリングすることは現実的ではない[6].

また，多くのパターン検出に関する先行研究は，クラスタ数やエラーの閾値等のパラメータ設定やチューニングが必要であり，これらのパラメータが出力結果に大きな影響を与える．したがって，時系列データからのパターン検出においては，解析する時系列データの事前知識を必要とせず，ユーザの介入を必要としない手法が望ましい[6].

上記のことから，関連研究 3-1 で述べた「深層学習を用いた時系列データの要約と分類」を文書生成の方法として採用する．

文書生成タスクにおける各項目を抽出する方法は様々であるが，今回は以下のツールを参考とした．

＜ツール＞

(1)診療録（プログレスノート）

(2)DWH（テンプレート）

(3)退院時要約
(4)双方向対話型人工知能による総合診療支援システム
（ホワイト・ジャック）「次世代地域医療データバンク」

＜ツールの内容＞

(1)入院時所見，入院後経過，手術情報，インフォームドコンセント等
(2)入院時所見，入院後臨床経過，インフォームドコンセント等
(3)入院までの経過，主訴・現病歴・既往歴，入院後臨床経過等
(4)診療情報，医療検査情報，医薬品情報，地域特性情報，気象情報

診療録 (プログレスノート･テンプレート等) に関しては，記載漏れが発生するという問題点がある．
この場合，記載漏れに関する指摘事項を推定する SVM を構築し，それを用いて注意を促すことが必要である．
上記のような問題が発生することを防ぐため，総合診療支援システム等で記載漏れを防ぐことが重要である．
いずれのツールに関しても，診療情報がカテゴリ別に分別されていない場合についての監査方法を以下に手順を述べる．

＜多次元時系列データの集合を以下の手順で作成する＞
形態素解析（文章の意味を解析）
↓
分散表現

【問題定義】
多次元時系列データの集合　D＝全ての患者情報（診療録）とする．各文書の監査項目及び内容については，{主訴・現病歴・既往歴入院時所見等のデータが含まれるもの} とする．
↑X1 がある一人の患者とする。患者はＸn 人いるとする。
↓
単一の時系列シーケンス Ｘ をｍ個のセグメント集合Ｓ＝｛ｓ1,・・・ｓｍ｝に分割

・例　「退院時要約（複数患者)」をデータセットとする．
＜時系列データからの自動特徴抽出＞
・レジーム 1
　入院までの経過
・レジーム 2
　主訴・現病歴・既往歴
・レジーム 3
　入院後臨床経過の記載
・レジーム 4
　退院時の処方

上記のような内容の時系列パターンを発見する。
↑各セグメントの重要度 a を求める
↓
記載のレベルを 5 段階とし，分類するラベル y は以下の 2 つとする．

レベル 2 以下　→　あまり記載のない記録
レベル 3 以上　→　概ね記載のある記録

5. 文章自動生成における主な手法と独自性に関する―考察 [12]

5.1 自動要約による文章自動生成 [12]

自動要約の古典的な H.P. Luhn は，テキスト中の 重要な文を抜き出し，それを出現順に並べることによってそのテキストを読むべきか否かを判定するといったスクリーニングのための要約が自動生成できること を示した．つまり，自動抄録に似ており，「理解し，再構成し，文章生成」というのではなく，「理解する箇所 が重要部に近似する」と割り切って考えたものである． 重要語の決定には，単語の頻度を用いるなど，現在の 自動要約の流れは，H.P. Luhn の影響が少なくない． また，ニューラルネットの文圧縮の研究も進んでおり，seq-to-seq モデルでは ROUGE スコアの低下はモデ ルへの入力文が長すぎると新聞記事のヘッドライン生 成が劣化する問題点がある．Attention の付いていない encoder-decoder model を使用し，encoder には片方向 LSTM を適用し，最適化には adam を用い，出力時には beam-search を用いるなどが良い結果が出ているとされている.さらに文抽出手法を強化学習にしたテキスト自動要約手法の研究も行われている[12].

5.2 手法の生成文の考察・データセットと各手法の詳細 [12]

・ 単語出現頻度に基づく文章要約

ここでは，H.P. Luhn(1958)による要約アルゴリズムを 基に簡略化したものを用いた[12].

1. 形態素に分解し，各段落で単語の一覧を作成する．
2. 段落内で、もっとも多くの単語を含む文を探し，ランキングにする．
3. ランキング順に表示する．

5.3 実験で用いた手法の長所・短所[12]

自動要約（頻出キーワード→それを含む文→昇順に並べ返す）
・メリット : 文と文とのつながりが不自然でないこと[12].

5.4 評価結果 [12]

・文章 2（自動要約 ） 5 点
（例文 ） 私の知り合いの老人 Y さんは現在 90 才の元

気な男性。 Y さんの健康法は毎日 2 時間 くらいは散歩を続ける事だ そうです。それも晴の日だけでなく、雨の日も散歩に行かれ ると言うのでびっくり。本人いわく「この年で仕事 もないので、私は散歩する事が仕事 と思って毎日歩いているので、雨の 日でも行きます。雨だから今日は仕事が 休みとは普通なら ないでしょう・・・」との事でした。流石に脱帽です。 実はこ んな事があったそうです。お 医者さんから「もう 90 才になる のだから、あまり無理して歩かないほうがよいですよ。」と言わ れ、Y さんも「そうかなー」と思い 1 ヶ月近 く散歩を止めて いました。そしたら、バス停から家までの道のり約 5 分くらい の 緩やかな坂道が、途中に一度休 まないと息が切れて歩 けなくなったそうです。それで「これではまずい!」 と思 って、また歩き始めて 3 週間くらい歩き続けたら元に戻ったそうで す。歩く事は健康の基本です。半身の静脈の 流れを良くし、 身体の基礎筋肉を維持し、心肺機能を維持する事ができる のです。また、腰痛の 70%はしっか り歩くだけでも改善され ています。現 代は飽 食による肝 脂 肪が増えています。私 も 最近は運動不足なので、昨年 の 10 月からは子供と毎月 1 回は山登 りをするようにしています。皆 さんも運 動不足 と思 われる方は是非散歩を お勧め致します。毎日 1 時間は歩 いてほしいですね（572 文字）

（実務者の評価 ）語句の使い方や文章 としてきわめて自然であり、前後の文脈もつながっている．この精度で文 章生成であれば二重丸[12].

6. 単一言語コーパスにおける文の自動対応付け手法［13］

6.1 依存構造木の経路とそれらの類似度に基づく対応付け [13]

従来法の問題点を解決するため，「単一言語コーパスにおける文の自動対応付け手法」ではＤＴＰを単語とその意味カテゴリを要素とするノードの系列データをしてとらえ，ＤＴＰ間の類似度を部分属性列の一致に基づき計算し，各ＤＴＰに対してもっとも類似するＤＴＰを有する文を対応文とする手法が提案された[13].

要約文のすべてのＤＴＰに対し，最も類似したＤＴＰを持つ文を対応文とすることで，複数の対応文を過不足なく割り当てることができる．また，ＤＴＰを文節列とはとらえず，単語と意味カテゴリを要素とするノード列であるとみなすことで，語彙の言い換えも含みつつ柔軟に類似度を計算できる．

6.2 文から DTP への分解 [13]

日本語は語順が柔軟であることを考えると（意味を保持したまま）語順が変化しても依存構造は変化しない場合が多いことから，文を依存構造木として表現することは適切である．しかし，文間の対応関係は，ある文の一部分に別の文の一部分が対応するといった形になっているので，依存構造間の類似度を直接計算し，対応文を決定する手法は適していない．

つまり，文を依存構造木で表現したうえで，その部分構造に対して対応文を割り当てることが必要となる．

6.3 DTP 間の類似度計算法 [13]

従来より，文，文書などの 2 つのテキスト（それぞれ，$o1, o2$とする）間の類似度には，式(1)で定義されるコサイン類似度（距離）を用いることが多い．

$$simcos(o1, o2) = \frac{\sum t\ Wt, o1 Wt, o2}{\sqrt{\sum t\ Wt^2, o1 \sum t\ Wt^2, o2}} \quad (1)$$

ここで，$wt, o1, wt, o2$ は，$o1, o2$に出現する単語 t に対する重みであり，TF，IDF，TF・IDF などが用いられる．これは，テキストに出現する個々の単語を独立に評価して類似度を計算しているといえる．しかし，重みを考慮したとしても単純に個々の単語の重なりを見るだけでは，単語の並びを考慮していないので，テキスト間の類似性をとらえきることはできない．このような問題を背景として近年では，単語の組合せ（共起）を考慮してテキスト間の類似度を計算する手法が提案されている．たとえば，機械翻訳結果の自動評価には n-gram に着目した手法，テキスト分類には，テキストを文字の系列データと解釈してその系列パターン（部分文字列）に着目したストリング・カーネル（SSK）や SSK を拡張したワード・シーケンス・カーネル（WSK）を用いた手法が提案されている．

本論文では上記 WSK を改良した拡張ストリング・カーネル（ESK）を用いて DTP の類似度計算に用いる[13].
以下にカーネル関数について説明し，SSK，WSK について説明する．最後に ESK の詳細を述べる．

(4)　カーネル関数（K）は，2 つの対象 x，x´に対し，それらをある関数で写像した空間においての内積計算として定義される．つまり，$\boldsymbol{\phi(x) \cdot \phi(x') = K(x,\ x')}$となる．ただし，K は，対象を陽に$\boldsymbol{\phi}$で写像することなくその値を効率的に計算できる．K が$\boldsymbol{\phi}$で写像した空間での対象間の内積であることを考えると，x と x´の類似度を表していると考えることができる．

SSK は，入力対象をテキストとし，$\boldsymbol{\phi}$ がその部分記号列を基底とする空間へ写像することに相当するカーネル関数である．このとき，基底に対する座標値は，着目した部分記号列の重み付き総和であり，重みはスキップも含めた部分記号列長lに応じて減衰パラメータを用い，λ^l で与えられる．特別な場合としてλ を 1 とすると，2 つの記号列間の SSK のカーネルの値は，両者に共通に含まれる部分記号列の数の和に相当する．たとえば，「abaca」と「abbab」という記号列に対して，3 個の組合せ（以降，組み合わせ

る記号の数を d と表す）からなる部分記号列を考えると，これらの記号列に共通で，かつ，記号のスキップを許したうえで 3 つの記号の組合せ（d＝3）からなる部分記号列は「aba」だけであり，それぞれに 2 回出現している．よって d＝3 のとき，カーネルの値は 4 となる．ここで，組合せ数 d だけの部分記号列のみを計算の対象としていること，スキップも含めた部分記号列の長さℓに依存して λ^{l} が部分記号列の重みとして与えられることを考えると，d が大きい場合や部分記号列が長い場合にはカーネルの値が小さな値をとる傾向になるという問題がある．

そこで WSK では，以下の拡張を SSK に対して施している．

- d 個以下のスキップを許したすべての部分記号列の組合せを考慮する．
- スキップを含めた部分記号列長ℓではなく，スキップした記号の数ℓ-d に応じて λ^{l-d} で部分記号列の重みを与える．

これは，i 個の記号組合せによる SSK を $K_{ssk(i)}$ として表すと，$K_{wsk(d)} = \sum_{i=1}^{d} K_{ssk(i)} / \lambda^{i}$ であることを示している．

ここで，対象「abaca」と「abbab」を入力として K_{wsk}('abaca', 'abbab')の値を計算する．両者を $\boldsymbol{\phi}$ で写像すると最大 3 個までの部分記号列となる．よって，$K_{wsk(3)}$ は以下の式で求まる．

$$k_{wsk(3)}(\text{'abaca'}, \text{'abbab'}) = \text{aba}\{(1+\lambda^2)\times 2\lambda$$
$$\text{ab}\{+1\times(2+\lambda+\lambda^3)$$
$$\text{aa}\{+(2\lambda+\lambda^3)\times\lambda^2$$
$$\text{ba}\{+(1+\lambda^2)\times(1+\lambda)$$
$$a\{3\times 2 + \text{b}\{1\times 3. \quad (2)$$

しかし，WSK は，表層での単語マッチングしか行わないので，元テキストを言い換えて要約を生成した場合にはうまくマッチングがとれない．そこで，単語の意味属性も考慮してマッチングをとるため，WSK に対して「記号」を単語とその意味カテゴリを要素とする「ノード」として扱えるように改良を加えたカーネル（ESK）を提案する．ESK を用いることで，言い換えを吸収したマッチングが可能となる．

「11 月の雨」というテキストはそれぞれの単語と意味カテゴリからなるノードとして表される．1 つの単語に対して複数の意味カテゴリが割り当てられていれば，ノードは単語と複数の意味カテゴリから構成される．ESK における写像関数 $\boldsymbol{\phi}$ は，テキストを単語または意味カテゴリの部分要素列，単語と意味カテゴリの双方を含む部分要素列を基底とする空間へと写像する．d = 3 とし，「11 月の雨」というテキストを $\boldsymbol{\phi}$ で写像し，その空間で内積を計算したものがカーネルの値となる．

厳密には，$K_{esk}(T,U)$は以下の式で定義される．ただし，T，U はノード列を表す．

$$k_{esk}(T,U) = \sum_{m=1}^{d} \sum_{ti\in T} \sum_{uj\in U} K_{m(ti,uj)}. \quad (3)$$

$$K_{m(ti,uj)} = \begin{cases} val(ti, uj) & \text{if } m = 1 \\ otherwise \\ \{K'_{m-1(ti,uj)} \cdot val(ti, uj) \end{cases} \quad (4)$$

$K'_{m}(ti, uj)$ は以下の式で定義される．ただし，ti ,uj はT,U のノードを表す．val(ti , uj) は，ノード間で一致した要素数を返す関数である．

$$K'_{m(ti,uj)} = \begin{cases} 0 \text{ if } j = 1 \\ otherwise \\ \{\lambda\ K'_{m(ti,uj-1)} + K''_{m}(ti, uj - 1) \end{cases} \quad (5)$$

さらに，$K''_{m}(ti, uj)$ は以下の式で定義される．

$$K''_{m(ti,uj)} = \begin{cases} 0 \text{ if } j = 1 \\ otherwise \\ \{\lambda\ K''_{m(ti-1,uj)} + K_{m}(ti - 1, uj) \end{cases} \quad (6)$$

なお，類似度としてカーネルの値を用いる場合には

以下の正規化を施した式を用いる．

$$Sim_{esk}(T,U) = \frac{Kesk(T,U)}{\sqrt{Kesk(T,T)Kesk(U,U)}} \quad (7)$$

同じ内容を表す異なるテキストが与えられた際に文間の対応付けを行う手法を提案した[13]．提案手法は，文を依存構造木における経路集合ととらえ，各経路に最も類似する経路を有する文を対応文とすること，経路を単語と意味カテゴリを要素として持つノード列として表現し，ESK を用いて類似度を計算することを特徴とする．TSC の単一文書要約データ，複数文書要約データに対し人間が対応付けを行った結果を正解として提案手法を評価した結果，単一文書要約のデータに関しては F 値 0.95～0.97，複数文要約データに関しては F 値 0.72～0.83 という高い性能を示し，従来の対応付け手法より優れていることを確認した[13].

7. 考察

・7.1 深層学習を用いた時系列データの要約と分類

既存手法に基づいて退院時要約記載に関する監査担当者の主な指摘事項を推定する自動監査システムを提案したが，このような手法を用いて将来的には全記録へ拡大することを目的としている．今後は，監査点検用紙に関する法令・診療報酬上の根拠などに該当する患者データを一つのデータベースへ蓄積し，そのデータを用いてこのようなシステムで要約・分類することで監査担当者の業務負担を軽減することが可能であると思われる．

・7.2 文章自動生成における主な手法と独自性に関する一考察

・7.3 単一言語コーパスにおける文の自動対応付け手法

先程説明した既存手法に基づき診療記録全体の文書自動生成が可能であると考える．あらかじめ医療従事者が診療録に記載したものであるならば文書生成可能である．しかし，元データとなる診療録に監査項目を満たす記載がない場合があるため，将来的には，記載漏れがある箇所にアラートがかかるようなシステムの構築が必要となってくると考える．

また，現病歴・既往歴等過去の記録については，他医療機関からの紹介状等のデータを一つのデータベースへ蓄積し，そこから監査項目に該当するデータを抽出し文書自動生成することが望ましいと思われる．上記に記載した過去の記録や入院後臨床経過等の時系列データに関しては，以下の項目を学習するシステムの構築が望ましいと考える．

- 病名等から時系列パターンを推定可能であり，過去の患者と同様又は類似した時系列データの場合，そのデータを学習し，最終的には過去の全てのパターンを学習する．
- 上記に該当した場合は，その時系列データを必要に応じて該当箇所へ自動で流し込む．
- 新たに必要となった情報を医療従事者が追記し，その情報をシステムが学習する．

このようなシステムを構築することで，医療従事者の記載負担を軽減することが可能であると考える．

参考文献

[1] 社会保険研究所．令和 2 年版保険医療機関のための診療報酬とカルテ記載
[2] 日本診療情報管理学会．診療情報学 p．167-168
[3] 田中　幸三．“医学管理料算定に関する課題とシミュレーション”．https://www.nec-nexs.com/supple/medical/column/tanaka/column066.html
[4] 岩穴口孝.継続診療に繋がる退院時要約作成支援システム開発のためのデータマイニング技術の応用．科学研究費助成事業研究成果報告書
[5] 三菱総合研究所．“実用化が始まる文書生成 AI 第２回：概要と企業信用調査レポート作成支援の検証”．https://www.mri.co.jp/knowledge/column/20200615-02.html
[6] 山室 冴，松原 靖子 他．深層学習を用いた時系列データの要約と分類．DEIM Forum 2018 C3-3
[7] 東山翔平，関和広，上原邦昭．医療用語資源の語彙拡張と診療情報抽出への応用．自然言語処理　June 2015，vol．22 No．2
[8] 村脇有吾，黒橋禎夫．形態論的制約を用いた未知語の自動獲得．言語処理学会 第 14 回年次大会 発表論文集(2008 年 3 月)
[9] 笹田鉄郎，森信介，河原達也．自動獲得した未知語の読み・文脈情報による仮名漢字変換．自然言語処理 June 2010，vol．17 No．4
[10] 美野英弥，伊藤均，後藤功雄，山田一郎，徳永健伸．ニューラル機械翻訳での目的言語側の文脈の効果的な利用．自然言語処理 vol．28 No.4．December2021.
[11] 深津博．電子カルテ代行入力入門．P78．経営書院
[12] 太田博三．文章自動生成における主な手法と独自性に関する一考察．人工知能学会研究会資料
[13] 平尾努，鈴木潤，磯崎秀樹，前田英作．単一言語コーパスにおける文の自動対応付け手法．情報処理学会論文誌．Oct.2005.　Vol.46 No.10