博士論文

自然言語処理を利用した本邦版 Computer Assisted Coding 構築手法の提案と性能評価

２０２３年

福井工業大学大学院　　工学研究科博士後期課程
応用理工学専攻

辻岡　和孝

目次

はじめに

筆者は、過去 30 年間、一貫して医療関係の仕事に従事している。福井県立三国高校普通科を卒業後、倉敷市にある川崎医療福祉大学医療技術学部医療情報学科に進学した。市原清志助教授（現：山口大学名誉教授）のゼミナールに配属し、2 年半にわたりプログラミングのご指導を頂いた。市原先生は医師でもあることから、一人暮らしで不健康な生活を送っていた私にいろいろと体調管理のアドバイスを頂いた。研究テーマは HRA という生活習慣病の危険度を予測するシステムの開発というテーマを与えていただき、4 年時には、神戸市で開催された第 18 回医療情報学連合大会において卒業研究の内容を発表した。大学卒業後、金沢市の石川コンピューター・センターに入社し、医事会計システムとオーダエントリシステムの開発と導入を経験した。その後、福井県済生会病院に転職し、医療情報システムの管理業務のみならず、臨床工学技士としての透析業務や医療機器の管理を経験した。同部署で開発した統合型医療機器管理システムは、福井医療株式会社（現：ミタス）へライセンス譲渡を行い、全国展開を実現することができた。福井県済生会病院医療情報課では主任として病院情報システムの導入と管理を行う事務的な業務ではあるものの、シンクライアントやアクティブディレクトリ等、当時としては先進的なアーキテクチャを積極的に採用し全国的にも注目を集めることができた。在職中に福井県立大学大学院のビジネススクールに通い、異業種でありながら共通の目標をもつ仲間とお互い刺激を与えながら経営学修士（MBA）の学位を取得した。修了後は富山大学学術研究部に助教として採用され、アカデミアの世界に踏み入れた。学会活動に精力的に取り組み、国内初となるスマートフォンタイプの電子カルテを導入したり、携帯端末ならではの新規機能を盛り込み導入前後での評価を行い、論文化した。また、クリニカルパスシステムの導入や、小型医療機器から効率よく電子カルテへ転送する仕組みを構築し評価した。その他、第 39 回医療情報学連合大会の事務局長を拝命し、全国規模での学会運営という貴重な経験をさせて頂いた。

アカデミア所属であることから軸となる研究テーマを模索していたところ、ちょうど富山大学に赴任した 2013 年頃より word2vec の登場もあり自然言語技術が飛躍的に発展してきた。色々調べていくうちに米国ではカルテ記事の内容から、病名を推測するシステムが既に存在することを知り、日本語においても同様なことができないかと実験したところ、予想以上の結果が得られ、研究速報として論文化した。自然言語処理に関して、さらに研究を深く進めるべく福井工業大学大学院工学研究部応用理工学専攻の研究博士後期課程へ夢と希望に満ち溢れる状態で進学したのが 2019 年の春である。

本論文の研究テーマである、「自然言語処理を利用した本邦版 Computer Assisted Coding 構築手法の提案と性能評価」は、工学的な技術論にとどまらず、診療録の知識や、病院における診療科毎の入力傾向の知識、しいては医療制度の変遷の理解等の総合的な知見がないと、評価結果の考察にはたどり着かない。考察では、筆者の総合的な知見を集約し、まとめた。

論文の核となる Computer Assisted Coding による病名推測の評価部分は、日本病院会が母体である診療情報管理学会誌「診療情報管理」にて原著論文として採択された [1)]。診療情報管理学会において、筆者は診療情報管理指導者を拝命しており、診療情報管理士のさらなる知識の向上のための学術活動をしている。前職の金城大学に引き続き現職である国立国際医療研究センターにおいても単に診療情報管理士の養成にとどまらず、学会のシンポジストや職能団体である診療情報管理士会の総務委員会ワーキングにも所属し精力的に全国に情報発信をしているところである。

昨年、多方面での識者の意見を取りまとめた次世代電子カルテシステムの提言には、電子カルテシステムのデータの利活用に関しても盛り込まれている。診療録に記載すべき項目の標準化の重要性、自然言語から機械可読性が可能な構造化されたデータに変換し蓄積する研究もスピード感も持って行われている。時代がドラスティックに変化している現在は、新しい仕事のスタンスに対応していけなければならない。本研究では、電子カルテの退院サマリーの情報から自動で病名のコーディング行う。これは従来、診療情報管理士の独占的な業務の一部を自動で行うものであり、診療情報管理士が定型的な業務から解放され、より生産性の高いデータアナリストやデータサイエンティストへ業務がシフトをしていくことを願い、実験をすすめてきた。本研究が論文という形で一般に公開され、診療情報管理士の業務範囲の拡大と地位の向上に繋がることを期待して本論文を作成した。

第 1 章

研究の背景と目的

1.1　研究の背景

国内では 1999 年に電子カルテが正式に認可され、20 年以上が経過した。現在、全国の 8 割以上の病院に電子カルテが導入されている。しかしデータを利活用するという意味では、ユーザフレンドリなインタフェースが提供されず、未だにシステムエンジニアによる SQL を用いたデータのエクスポートが必要である。システムエンジニアは必ずしも医療や病院業界に造詣が深いとは限らなく、医師や経営層が求めるデータを取得できない場合も多い。そのため、従来、紙の診療録の管理業務をしていた診療情報管理士にデータアナリストやデータサイエンティスト的な業務への拡張が期待されてきている。診療情報管理士は医学知識を持ち合わせており、従来より診療録の内容を統計処理し年報を作成したりしてきたが、DPC/DPPS という新しい診療報酬制度下では、病院の収益に直結する国際統計分類に基づいた病名をコード化する専門職として広く認知されてきている。

令和 4 年度診療報酬改定では診療録管理体制加算の要件として従来は専属の診療情報管理士の配置を求めていたところであるが、新たに情報セキュリティ教育や次世代医療規格である HL7 FHIR への対応状況の報告が求められた。HL7 FHIR は、REST API を用いた医療情報交換規約である。診療情報管理士にとっても従来の紙ベースでの管理から電子ベースの管理への業務のパラダイムシフトが求められてきているところである。

本邦においては診療情報管理士が、データアナリストとして活躍することはあっても、データサイエンティストとしての事例はまだ少数である。海外に目を向けると米国診療情報管理協会が実施した診療情報管理士が今後必要になるスキルについてのアンケート結果（図 1.1）によれば、ビッグデータアナリシスや情報学、データマイニングが上位にあり、レコード管理業務やコーディング業務のランキングは下位となっている [2)]。

米国では Computer-Assisted Coding（以下、CAC）と呼ばれる医療文章に自然言語処理を適用することで得られる機械学習モデルを導入した自動コーディングシステムが既に普及され、診療情報管理士の生産性を向上させていることが、レコード管理業務やコーディング業務のランキングを押し下げている要因と思われる。

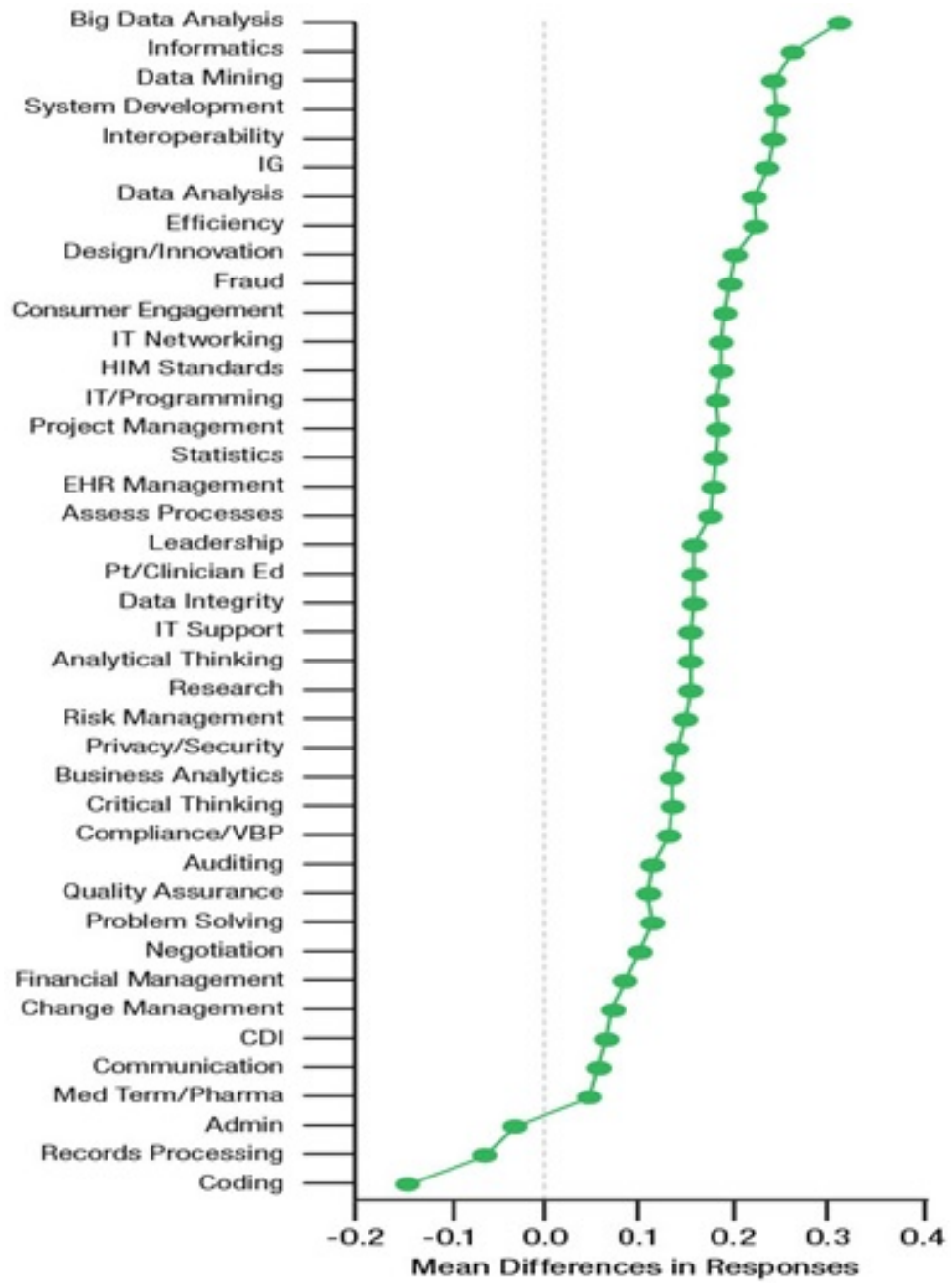


図 1.1　米国診療情報管理協会が実施した診療情報管理士が今後必要になるスキルについてのアンケート結果

1.2　先行研究

1.2.1　海外事例

CAC を開発している 3M 社のホワイトペーパーによると [3]、診療情報管理士が自然言語処理という新しい技術を受け入れることで、ワークフローを合理化するとある。米国では CAC は診療情報管理士の業務を奪う脅威な存在ではなく、導入することで作業効率を高めるツールとして受け入れられている。また Campbell[4] らは CAC の限界について調査し、CAC の標準規格がない点を指摘した。標準規格のためにはデータの完全性が必要で、完全でないデータからは CAC エンジンが用語の概念がない場合、エンジンはその用語を認識できず、誤ったコードを割り当てるかもしれないと問題提起した。

Fraser[5] らは退院サマリから病名や症状、医薬品名や治療行為、検査値や検査名を意味する医療表現を抽出するタスクにおいて Clinical BERT などの医療テキストから事前学習した BERT を使用することで、従来の分散表現と比較して高い精度を得られることを明らかにした。また同様のタスクにおいて、Wikipedia などの一般的な事柄について

記載したテキストを事前学習に用いた BERT と Clinical BERT を比較した場合、後者がより高い精度を示したことを報告し、対象タスクと同じドメインのテキストで事前学習を行うことの重要性を示唆した。

1.2.2　国内事例

本邦では、医療文書に関する自然言語解析は、大学病院とコネクションをもつ研究者が中心となり、進められている。鈴木ら [6] や野口ら [8]、木村ら [9] により退院サマリーの内容を自然言語処理にて解析し、病名を推定する研究がなされている。鈴木らは複数医療機関の退院サマリーのテキスト情報を形態素解析と、重みづけとして tf × idf 法を用いベクトル空間モデルを構築した。野口らは IRIS NLP を用いて退院サマリーの「主訴」と「既往歴」からエンティティを抽出し、診断群分類毎の特徴語を構造化する試みをしている。木村らは、DPC コードと関連の深いキーワードとの 2 次元の関連表を作成し、各機械学習手法の評価を行った。

退院サマリー内の非構造化テキストで入力される経過要約から病名を推定するためには、病名毎に特徴量を抽出し分析する必要があり、医療文書から病名等の固有表現の抽出には検証のための大規模な医療文書コーパスが必要となる。

鈴木は、研究機関である大学病院を中心に全国の退院サマリーと診療報酬データの病名コードである DPC コードを収集し、それぞれの大学の退院サマリーで学習モデルの評価を行った [7]。鈴木は目的変数として上位 20 以上の DPC コードとしている。結果を図 1.2 に示す。

検証＼モデル	千葉大学	香川大学	高知大学	長崎大学	大阪大学	聖路加	統合データ
千葉大学	68%	39%	53%	47%	54%	46%	67%
香川大学	33%	55%	41%	29%	39%	31%	52%
高知大学	43%	47%	69%	57%	60%	53%	70%
長崎大学	43%	41%	55%	67%	55%	47%	68%
大阪大学	51%	44%	59%	51%	72%	52%	70%
聖路加	52%	41%	52%	50%	55%	67%	65%

図 1.2　上位 20 以上の DPC コードでのクロスマッチ結果

鈴木は考察で、「DPC クロスマッチ自動判定の結果では、以前と同様に自施設のモデルデータでは高く、他施設のデータでは 10〜20 %程度低下するものの、全施設統合デー

タでは自施設と同等の判定率を示した。」と述べている。

鈴木らと同様の研究をすることは容易ではない。診療録には病名などの要配慮個人情報が含まれる。個人情報の取り扱いが厳密であることから、特に研究機関でない産業界からの進出には、情報を保持している機関との共同研究という形を取ったうえで、厳密に情報管理をする必要があり、研究を始めるに際しての倫理審査委員会や同意取得等の手続きの複雑さから、研究自体がスタートできない場合もある。診療録の情報管理が厳密であることから、大学のような学術研究機関以外の施設からの参画に障壁があり、企業が簡単に実験ができないこともCACの開発が進まない要因と思われる。

米国では診療記録は自然言語による自由記載が本邦ほど許容されていない。曖昧な表現による訴訟時のリスク回避の意味もありテンプレート等で各項目が定義される。また英文は自然言語部分も単語間がスペースで区切られているため、形態素解析において品詞の判断がしやすく自然言語処理がしやすくCAC構築を行いやすい環境といえる。しかし本邦では経過記録部分はワープロ入力による自由記載であること、日本語は単語が連続しているため形態素解析の難易度が高い事や、略語、例えば”プロポフォール”を”プロポ”と記載したり表記揺れが大きいことが国内でのCAC開発の障壁となっていると考えられる。その他、コーディングの専門家である診療情報管理士の養成が遅れたことから、正確なコーディング自体ができていなく、教師データが乏しい事もCAC開発が進まない要因と思われた。

1.3　研究の目的

2019年に標準的な退院サマリーとなる「HL7 CDAに基づく退院時サマリー規約」が厚生労働省より認定された[10)]。標準化された退院サマリーのフォーマットでは入院日、生年月日、病名等は必須項目であり、病名は標準病名マスターより選択されICD10コードが付与される。退院サマリー部分には「主訴」、「入院までの経過」、「入院経過」、「退院時状況」、「退院時使用薬剤情報」等が格納され、自然言語形式で保存することとなっている[11)]。

本研究では、本邦におけるCAC構築手法の提案を行う。退院サマリーの経過要約に自然言語処理を行い、「264種類の特徴単語毎のベクトル値」を求める。「264種類の特徴単語毎のベクトル値」と標準化された退院サマリーの入力項目でもある「年齢」と「性別」、また標準化された退院サマリーの必須入力項目の「記載者情報」と関連が深い項目である「診療科名」を説明変数とし、「診断病名コード」を目的変数とする教師あり機械学習を行う。

機械学習に使用する退院サマリーは、異なった電子カルテシステムから「診断病名コー

ド」、「年齢」、「性別」、「診療科名」、「経過要約」をそれぞれ取得する。電子カルテシステム変更になった条件下でも、記載項目が統一されていれば、CAC としての性能を発揮するかを検証し考察を加える。

本研究の貢献は、意味表現学習を採用し、医療文書をニューラルネットワークにより特徴単語毎のベクトル値に変換することで、結果を人間が理解しやすい形で表現することが可能となり、判定結果の解釈性の向上が期待できる構築手法を提案したこと。また、学習時の説明変数の設定を可能にし、混合行列の結果より診療情報管理士の知見を活かした評価を可能にした点である。さらに 2 種類の電子カルテシステムの退院サマリーデータを用い機械学習モデルの汎用的な評価が行えるようなベンチマークの手法を確立した。

鈴木らのように複数の研究機関からの退院サマリーの収集には、それぞれの研究機関での倫理審査委員会の了承を得る必要があり、研究を始めるにあたり制約が大きい。今回は、富山大学附属病院で稼働していた旧電子カルテシステム富士通製 NeoChart で作成した退院サマリーデータと、新電子カルテシステム富士通製 EGMAIN-GX Enterprise Edition の退院サマリーデータを用意することで、単一施設ではあるが疑似的に 2 施設の退院サマリーデータを収集したものとし評価する。また、診療報酬データは収集できなかったため、診療報酬データの病名コード（DPC コード）ではなく、診断病名コード（ICD10）で評価する。この 2 種類の退院サマリーのデータを用い、頻出する上位 20 位までの診断病名コードを調査し、その診断病名コードに対し、旧電子カルテ退院サマリーデータで作成した学習モデル、新旧両方の電子カルテの退院サマリーデータを使って作成した学習モデル、新電子カルテの退院サマリーデータを使って作成した学習モデルで性能評価し、考察を加える。

第 2 章

診療録を取り巻く環境

本論を理解するにあたり、医療分野に関して、用語の整理しておく必要がある。この章では電子カルテ、退院サマリー、診療情報管理士、クリニカルパス、CAC に関して解説する。

2.1　電子カルテシステム

医療機関の電子化は 1970 年代の医事会計システムと検査システムに端を発した。1980 年代後半には東京大学で初めてオーダエントリシステムが導入され、その後、全国に展開した。1995 年の Windows95 の登場により、Windows 上で動くオーダエントリシステムが急速に普及した。

1999 年に厚生労働省より「診療録等の電子媒体による保存について」が通達として出され、電子カルテの 3 原則（真正性、見読性、保存性）の条件下で、診療録を電子的に取り扱うことができるようになった。これが電子カルテの誕生である。当初は導入に躊躇している病院が大半であったが、2001 年には、「保健医療分野の情報化にむけてのガイドライン」が策定され、400 病床以上の病院の 6 割に電子カルテを導入するという目標が掲げられ、政府による補助金の後押しもあり着実に導入病院が増えていった [12)]。

今述べたように現在の電子カルテシステムはオーダエントリシステムから派生して開発された経緯があり、設計が医事への紙伝票をそのまま電子化したような仕様になっている。データの利活用まで想定したデータベース構造とはなっておらず、ベンダー毎にデータベース設計思想も違っている現状がある。データが大量にあっても、施設ごとにデータベース構造も違っているため、連係することが難しい状況が続いている。また病院側にも連携の必要性が乏しいこともあり、標準化という点、さながらデータを利活用する点において、米国に先を越されている。

海外の電子カルテは自然言語による自由文はなるべく使わないという風習があり、テンプレート形式での電子カルテが普及しているが、国内の電子カルテは自然言語で記載する割合が多くある。このため、記載内容を NLP を用いて機械的に解釈しようとした場合、表記ゆれや専門用語の複雑性から、海外の電子カルテよりも解釈の難易度が高いと思われる。

2.2　退院サマリー

診療録の記載は、医師法第 24 条第１項に「医師は診療したときは、遅滞なく診療に関する事項を診療録に記載しなければならない。」とあり、法的に義務づけられてる。診療録の必要事項は、医師法施行規則第 23 条及び療養担当規則第 22 条・保険診療における診療録の様式第１号で規定されており、保険医は、傷病名、診療開始年月日、終了年月日、主要症状、経過、手術及び処置等を記載しなければならない。診療録のフォーマットは指定されており、１号用紙、２号用紙、３号用紙に分類される。ただし診療録は通常外来と入院で分けることが通念であり、入院診療録の場合に、別途退院サマリーを記録する。

退院サマリーとは、入院中における患者の容体や経過、施した処置や手術、投薬や検査内容を要約し、A4 用紙 1 枚程度にまとめたものである。医療法においては診療録の一部とされる。自由文での記載され、かつ、術式や医療用語は、業界特有の専門用語が多用されているという特徴をもつ。診療報酬上では、患者が退院後 2 週間以内に退院サマリーを完成させる必要がある。仮に退院サマリーの完成率が施設内で 100 %を維持できなかった場合は、診療報酬上での減収されるというペナルティのほか、施設基準が満たせなくなることにつながり、しいては施設の機能評価係数の減につながる。そのため、診療情報管理士は退院サマリーがスムーズに完成するよう、医師事務作業補助者と連携し様々なサポート体制を敷いている病院が多くある。

退院サマリーに関しては従来、定められたフォーマットは存在しなかった。標準化の観点から、2019 年に日本診療情報管理学会と日本医療情報学会が合同で提唱した退院サマリー標準フォーマットは、HELICS 協議会を通じて、厚生労働省標準様式となった。現在は厚生労働省研究中山班にて医療情報交換規約である HL7 FHIR とのマッピング作業が進行中である。

2.3　診療情報管理士

診療録を管理する専門職として「診療情報管理士」という職種がある。診療情報管理士は一般社団法人日本病院会を母体とした診療情報管理学会という学術団体が認定する民間資格であり、2022 年現在、全国で 4 万人以上の認定者が存在する。上位の資格としては「診療情報管理士指導者」がある。指導者には現在診療情報管理士で活躍しているスタッフ向けに時代に即した教育指導をすることが期待されている。申請に必要な条件として、経常的な学会活動や論文執筆等が求められる。全国で認定者は 100 名程度しか

いなく、これは全体の 1 %に満たない。筆者は 2001 年に診療情報管理士の認定を受け、2020 年に診療情報管理士指導者の認定を受けている。

診療情報管理士は 2000 年以降に認定者が急激に増加した。これは日本の医療制度が DPC/PDPS と呼ばれる診療報酬制度に移行したことが大きく影響している。DPC/PDPS とは、病気毎に入院の 1 日当たりの診療報酬点数が決定される制度である。病名は医師が決定するが、DPC/PDPS で診療報酬請求をするためには、その病名を適切に標準コードに変換することが必要となる。診療情報管理士は、患者が入院に至った背景や治療内容、医師が記載した病名等を参考に、主病名や副傷病名を決定する。その決定した病名にコード付けを行う。このコード付けの作業をコーディングといい、診療情報管理士の業務の軸となるものである。DPC/PDPS では病名コードと入院中の患者の容体を経過記録の内容から判断し、診療情報管理士はその患者の DPC コードを決定する。DPC コードによって診療報酬点数が大きく左右することから、診療情報管理士は病院を経営していく上でも重要な役割となる。

近年は従来のコーディングやカルテ管理の業務から、DPC データを分析し、上層部に経営改善の提案をするような経営管理的な役割に業務拡張してきた。医師は診療録に病名を付するが、必ずしも、万人に理解できる病名ばかりではない。病名も手書きで記載することも多いことから、診療情報管理士は正確に国際的に標準化された病名に変換し、病歴を管理している。この国際的に標準化された病名集の事を国際疾病分類といいICD10(International Statistical Classification of Diseases and Related Health Problems ver10) と呼ばれる。世界保健機関憲章に基づいて、WHO が作成した。現在 ICD11 が最新のバージョンとなるが、国内導入は数年後のため、実質的には依然 ICD10 が利用されている。

ICD10 では 14,000 種類の疾病が分類されコード化されている。本邦においては感染症等の稀な疾患もコード化の対象となっている。例えば、新型コロナウィルス肺炎（Covid-19）は、当初は当てはまるコードはなかったため、エマージェンシーコード U07.1「 2019-nCoV acuterespiratory disease」」を利用していたが、その後 WHO より通達があり、B34.2 が割り当てられた。

ICD10 は死因統計に利用する目的で開発され、数ある登録病名から主病名を特定して、死亡診断書に書かれた病名に ICD10 をコード化する事から始まったが、医学的分類に応用が利くことから 2000 年以降に国内導入された DPC/PDPS 制度においては、最も医療資源を投入した病名に応じて、診療報酬点数が決定するようになり、ICD10 のコーディング能力が病院経営において大変に重要な位置を占めるようになった。

大規模病院になるほど診療上管理士は ICD10 コーディング以外の業務、例えばカルテ開示業務や委員会の事務局等の付随的な業務に追われるようになる。このため、コー

ディングなどの定型業務を AI や RPA で効率化することが、求められるようになってきた[13]。

2.4　退院サマリーの質的点検

退院サマリーの経過記録、SOAP 形式という記載ルールがある。これは、Dr.WEED が提唱した POS（Problem Oriented System）の概念から派生したものである[14]。S が Subject、O が Object,A が Assesment、P が Plan の意味となる。言い換えると、S には主訴として患者の訴えや医師からみた患者の様子を記載する、O には検査結果などの客観的データを記載する。A は S と O の結果から、求められる評価を記載する。P は最終的にどのような処置を行うかを記載する。これらが正しい形式で退院サマリーが記載されているかを診療情報管理士はすべての退院サマリーにおいてチェックを行っている。なお、退院サマリーの記載内容に不備がある場合は、差し戻すことはないが、監査という形で記載した医師にフィードバックする運用をしている病院が多い。

2.5　クリニカルパスシステム

クリニカルパスシステムは、1985 年にアメリカの看護師であるカレン・ザンダーがニューイングランド・メディカルセンター病院で提唱したものである。クリニカルパスシステムは、医療の質と効率を高めるための手法として医療に導入された。患者の入院から検査、手術、退院まで、疾患ごとに作成されたタイムスケジュールに従って治療を実施する方法である。日々、患者が計画通りに経過しているかどうかをアウトカム評価を行い、計画通りであれば計画を継続する。問題が発生した場合は、異変として取り上げられ、パス発生としてクリニカルパスシステムを終了するかどうかを判断する。

カレン・ザンダーは、チーム医療の重要性を再認識し、職種間の情報共有の重要性を訴えた。これにより、入院期間が短縮され、病床の回転率の向上が期待できる。さらに、治療の進捗状況を事前に把握することで、患者の不安も軽減される。さらに、必要かつ十分な検査や治療が行えるようになり、医療事故の防止や医療の標準化による医療の質の向上にもつながると指摘した。

日本の医療現場でもクリニカルパスシステムの有効性が認められ、主要な治療や検査で導入が進んでいるが、クリニカルパスシステムは、医療側からのアプローチだけでなく、病院経営に精通した診療情報管理士の関与により、病院経営側からのアプローチも可能となる。DPC/PDPS のもとでは、クリニカルパスシステムの導入は最終的に在院日数の短縮に寄与する。DPC/PDPS の制度上では大まかに入院日数が短いほど一日あた

りの診療報酬点数が高い。また、その診療報酬点数はICD10がベースとなるDPCコードに応じて変化する。このためICD10コーディングが重要になってきており、病名コードであるICD10と病名に応じて診療計画を立てるクリニカルパスシステムは親和性が高い。在院日数の長さが問題視されてきた日本においては、電子カルテに電子クリニカルパスシステムが実装されるようになった。筆者は臨床工学技士と診療情報管理士としてのライセンスを生かし、医療経営と医学の視点から富山大学附属病院の新電子カルテの導入を通じて、富山大学附属病院では初めて電子クリニカルパスシステム（図2.1）を導入した[15)]。　急性期を担う大学病院においては患者の容体が刻々と変化し、クリニカル

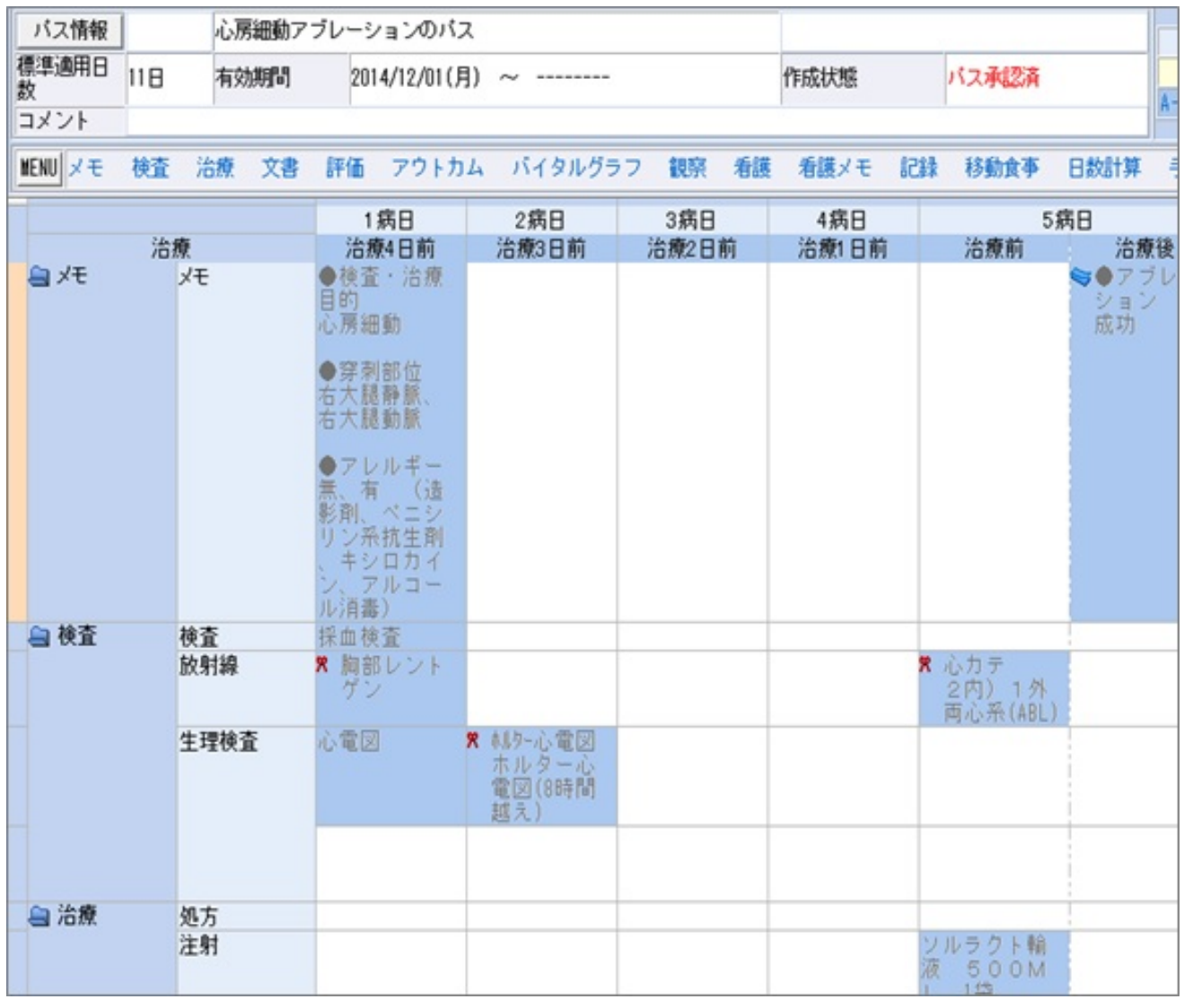


図2.1　富山大学の電子クリニカルパスシステム

パスシステムに即した計画的な退院が難しいことが判明した。稼働1年後の評価では9つの診療科で運用されていることが確認できた。入院を担う診療科が22の診療科であったため、4割の診療科で利用された（表2.1）。

稼働後1年後の退院患者数は、のべ10,964名であり、その中でクリニカルパスシステムの適用患者は1,412名であった。全入院患者の約13％は電子カルテでのクリニカルパスシステム機能を用いて処置が行われたことになる（図2.2）。

表 2.1　診療科別パス適応件数の詳細

診療科名	H27.1	H27.2	H27.3	H27.4	H27.5	H27.6	H27.7	H27.8	H27.9	H27.10	H27.11	H27.12	総適応数
第二内科	16	78	96	83	56	87	85	79	65	66	77	61	849
第三内科		8	4	4	1					17			
皮膚科									1	3	1	2	7
小児科												2	2
第一外科				1	3	3	1	6	1	9	18	19	61
第二外科		2	2	11	6	9	8	5	7	12	11	8	81
産科婦人科		1	2	3	4	3	3	12	19	26	40	36	149
耳鼻咽喉科	7	16	20	22	13	33	25	24	21	13	20	16	230
歯科口腔外科										3	6	7	16
月総数	23	105	124	83	125	172	122	126	114	132	173	151	1412

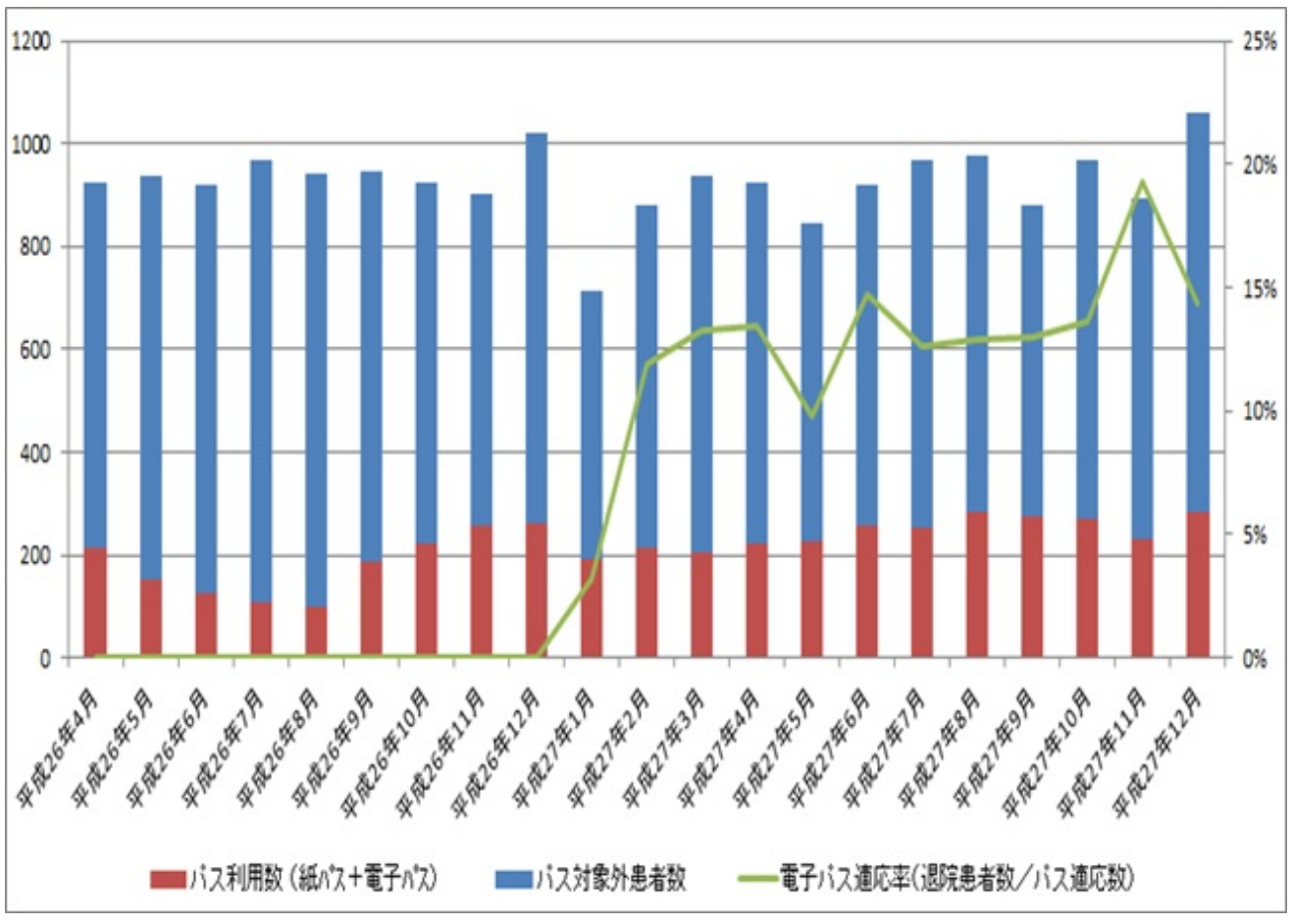


図 2.2　パス利用者数と電子パス適応率の推移

さらに日々のアウトカム評価の入力率の調査をした。アウトカム評価は日々患者の容体がクリニカルパスシステムの計画上想定内に入っているかを評価する。バリアンスは観察項目が想定外の場合に、関連する要因を入力するものであり、アウトカムが正しく入力されていることが前提とあるが、稼働後一年後の評価では、アウトカムの入力率が 26 %しかなく（図 2.3、図 2.4)、このことからクリニカルパスシステムを単なるオーダセットの機能という認識で利用している問題点が明らかとなった。

この様な状況を踏まえ、稼働 2 年目からはクリニカルパスシステムには入院中の一連の行為や評価が入力されるため、これらをまとめることで、退院サマリーを半自動的に生成する事ができると考え、クリニカルパスシステムの入力内容から退院サマリーを自動生成する機能を日本クリニカルパス学会で提唱した [17, 18]。

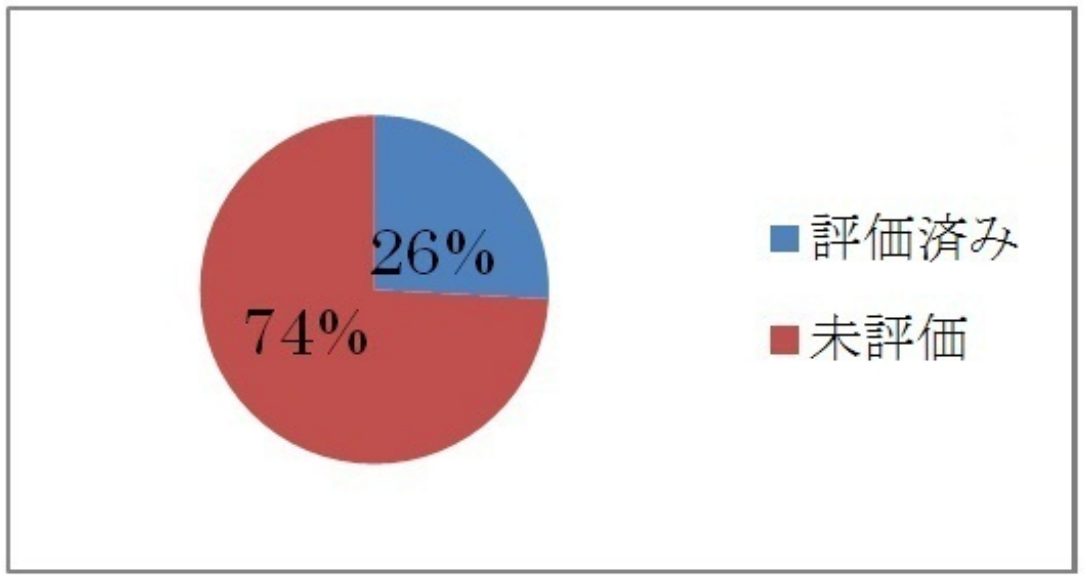


図 2.3　電子クリニカルパスのアウトカム評価率

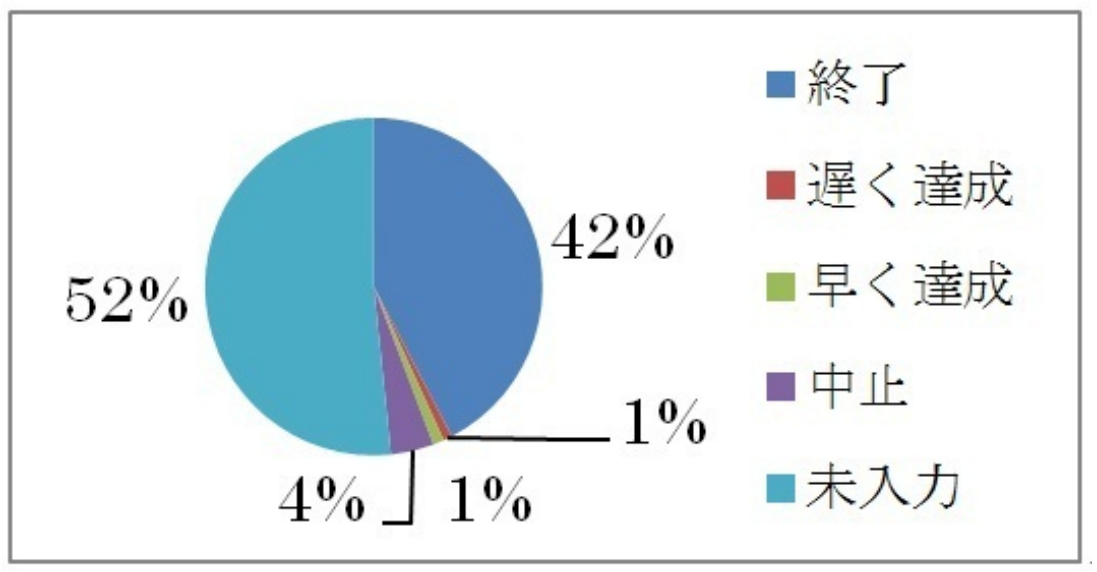


図 2.4　電子クリニカルパスのアウトカム評価内訳

第 3 章

国際疾病分類

CAC で求められた結果を定量的・定性的に評価し、考察していく上では、国際疾病分類についての知識は欠かせない。診療情報管理士が行う国際疾病分類のコーディングルールをすべて解説すると膨大になるため、この章では関連項目の概要のみ解説する。詳細が知りたい場合は厚生労働統計協会より発刊されている「疾病、傷害及び死因の統計分類提要　ICD-10（2013 年版）準拠」の第一巻（内容例示表)、第二巻（総論)、第三巻（索引表）及び日本病院会共済会より発行されている診療情報管理テキストを参照のこと。

3.1　疾病分類とその意義

ICD10 をはじめとする国際統計分類の体系は、国際的な比較から、各医療機関の診療内容までを評価するための、世界共通の標準化された規格である。医療機関で作られる「診療記録」と総称される情報には、各々の疾病の診断や処置、その他の医療行為などに関連する情報が含まれている。こうした個別の情報を、その診療科、医療機関、地域、さらには国レベルで収集・集計し、その集計地を他の集計地と比較することで、自らの属する集団の評価が可能となる。特に国のレベルでは、その評価は国際比較という形で、しばしば政策に反映される重要な情報となる。このために各国が比較評価可能な形で集計分類することが必要となり、WHO は、各国が表す疾病や障害および死因統計を、共通の分類方法を用いて国際比較、活用できるように「国際疾病分類」を制定し、これに沿って諸統計を作成するように勧告している。この WHO が定めた「国際疾病分類」に準拠して、わが国の特有の事情に合わせた修正を加え「疾病・障害及び死因の統計分類提要」として国内で使いやすい形で提供している。

3.2　国際疾病分類の使用目的

国は保健や福祉行政の企画、人口問題研究、医学研究に必要な材料として、各医療機関から集めた医療情報を用いて、死因統計や疾病統計を作成している。その結果を正確で信頼性の高いものとするためには、情報源となる死因や疾病の分類項目の内容基準を、

予め正確に定めておくことが大切である。また、疾病及び死因を国際比較、活用する場合、各国が共通した情報収集法に従って分類することが望ましい。それには、分類・統合ルールを医療機関が共通して採用し、同一方法で管理しておくことも重要な条件となる。

本邦でも WHO の勧告に従って、ICD10 を刊行している。今では、国の行政の情報だけではなく、各医療機関における疾病名のインデックスとして、また医療の監査やサービスの評価、病院経営・管理情報等の一手段として利用する病院も増えている。

ICD は、10 回にわたる大きな改訂を繰り返して作られてきた分類である。分類はいわゆる病名集ではなく、様々な病態・状態などを意味ある分類項目に当てはまるように工夫されたものである。例えば病名集に、ICD10 にあるような「その他の肺炎」という病名は存在しない。なぜなら「その他の」というためにはそれ以前に「A という肺炎」、「B という肺炎」のように「... という肺炎」が網羅されていなければ、それ以外の肺炎という状態が定義できないためである。分類では、その分け方に従って、この A、B にあたる肺炎がきちんと定義されこの A,B にあたる肺炎がきちんと示され、そのうえで、A にも B にも当てはまらないけれども、肺炎という病態であるということがはっきりしていれば「その他の肺炎」と分類できる、という立場に立つ。どのような状態であっても、ほとんどが ICD10 では何らかの分類に当てはめることができ、コード化ができる。コード化することで、同じ分類に当てはまるものは同じコードをつけることができる。患者 A と患者 B がほぼ同じ状態であることを表すことが可能となる。単純に病名だけでは、同じ病態かどうかの判断はつかないが、コード化することで、判断の基準が同一の病態として認識される。

3.3 国際疾病分類の構造

国際疾病分類のコードは、アルファベットと数字で構成されている。コードによって疾病や傷害の部位、原因などを表すことができる。最初のアルファベット 1 桁と数字 2 桁で疾病を表現する。その後ドット「.」が入り詳細分類として 1〜2 桁が割り当てられる。

近年は精度の高いコーディングが求められ、カルテの内容を読み解き詳細分類までしっかりとコーディングすることが求められている。以前は詳細不明コード「.9（ドットナイン）」が横行していた時期もあるが、現在は病院から提出された DPC レセプトコードにおいて、10 %以上「.9（ドットナイン）」があると病院自体がペナルティを受ける制度ができた。このため詳細分類まで正確にコーディングできる診療情報管理士が求められている。

なお、厚生労働省では原則毎年「ICD の ABC」[16] を発行し、国際疾病分類の簡単な解説書を発行している。図 3.1 は「ICD の ABC」で紹介されたコーディング例である。

図 3.1　厚生労働省「ICD の ABC」より一部抜粋

また、国際疾病分類はもともと死因統計であり、正確には、副傷病名も記載する必要がある。またどういった経緯で、死因の病名となったかの経緯がわかるような病名に関してもコード付けすることが求められている。そのため、ダブルコーディングの仕組みが設けられているが、こちらは死亡診断書に記載するときに求めれている。ただし、DPC/PDPS 制度下の診療報酬上のコード付けではそこまでは求められていない。現時点では DPC/PDPS で重要な病名は「最も医療資源を投入した傷病名」と「副傷病名」で

ある。DPC は厚生労働省が医療費が増加することを見据えて、病気毎に投入できる医療資源に制限を設ける意味合いで開発されたが、開発時点では有効な病名マスタが無かったため、死因コードとして既に存在していた ICD10 を流用したという経緯がある。そのため、実際に運用してみると、医師が想定する病名が ICD10 には存在しないこともあり、その問題を埋めるため、医療情報システム開発センターでは、医師が想定する病名とレセプト病名や ICD10 コードを紐づける標準病名マスターを開発し、厚生労働省標準規格としてリリースしている。

電子カルテシステムはこの標準病名マスターを病名登録機能に組み込んでいる場合が多い。既に普及しているが、そもそも紐づけているコードに誤りが指摘されたり、指定難病でありながら、未だに登録されていない病名もある等、未だ発展途上のマスターといえる。

3.4 国際疾病分類の自動化における取組み

診療録に記載された内容を各種統計や、ディープラーニングや機械学習等の手法を用いて解析し、候補となる ICD10 病名を提示するシステムのことを、CAC と呼ぶ。米国では既に普及が進んでいるが、本邦においては、未だ研究開発段階である。

本邦において普及を阻害している原因として、医療分野での AI 開発者が少ない点、退院サマリーのフォーマットが各社ばらばらでデータの統一性がない点、記載への自由度が高く解析が困難な点があげられる。本邦でも CAC が普及することで、診療情報管理士は労働集約的なコーディング業務に費やす労働時間が軽減し、浮いた時間で生産性の高い業務へ移行することが期待できる。前述したように退院サマリーの項目が標準化されたため、CAC の研究開発は幾分かはやりやすくなってきている素地はあるが、解決すべき課題は多くある。例えば、ディープラーニングは音声や画像認識の領域では先行的に研究が進んだものの、文字や単語といった記号を扱う自然言語処理の分野では、記号間の類似度や関連性を記号それ自体が持つ属性から直接計算することができないため、音声認識や画像認識のように簡単にはいかない。文書分類の問題を考えてみると、例えば診療録に「両側の唾液腺に腫脹が見られ、流行性耳下腺炎の疑いを認める」という記述があったとする。この記述を「両方の耳下腺に腫れが見られ、おたふくかぜではないかと疑う」と同じカテゴリに分類できるであろうか。それを可能にするには、「流行性耳下腺炎」と「おたふくかぜ」が同義語であることを知っていなければならない。その解決策として、シソーラスの利用が考えられるが、シソーラスの開発には多大な労力がかかるだけでなく、人間が作る以上、すべての用語をカバーすることはできないこと、仮にできたとしても、どう活用するかという課題が残っていた。

課題の克服が期待されている手法として、2018 年 10 月に Google より公表された

BERT[19] がある。例えば「疼痛」に代表されるような「痛」を含む表現だけではなく、「ズキズキ」や「ジンジン」など擬音語によって表現されることもあり、表現のバリエーションが広いことが問題が、BERT では Masked Language Model と Next Sentence Prediction というラベルなし事前学習ののち、事前学習で得た重みを初期値としてラベルありデータでファインチューニングを行うことで、自然言語を文脈レベルで理解することに成功した。

自然言語処理の分野は近年目覚ましい発展を遂げてきており、医療文書コーパスが用意できれば、CAC の構築において様々な手法で評価できる開発環境が整ってきている。

第 4 章

機械学習

本研究においては CAC の構築に機械学習を採用している。機械学習には様々な種類がある。この章では、各機械学習の特徴について記載する。

4.1 教師あり機械学習と教師なし機械学習

機械学習とは、多くのデータをプログラムで読み込ませ、大量のデータから出現パターンを導き出しデータの分類や予測を行うことである。大きくは、「教師あり機械学習」と「教師なし機械学習」に分けられる。

4.1.1 教師あり機械学習

教師あり機械学習とは、正解のデータをプログラムに与えながら、学習をさせていく手法である。正解データが用意されているため、予測結果の検証ができるため、正誤判定が明確な目的に利用しやすい。例えば、大量のデータがあるとして、それぞれの個々のデータに正解データをタグ付けしておくことで、プログラムがデータを読み込んだ時に、その正解データの特徴を少しづつ学習しておくことができる。

4.1.2 教師なし機械学習

教師なし機械学習は、学習データに正解が付与されていない学習方法である。教師あり機械学習では、正解のタグをベースに機械学習分類を行うが、教師なし機械学習では正解データが無いため、教師なし機械学習ではクラスタリングという手法を用いて学習していく。クラスタリングとはデータ分類方法の一つで、データの特徴を導きながらグルーピングを行っていく手法のことである。

4.2 機械学習の様々な手法

大量のデータから独創性のあるデータ解釈ができるようになった。その出現パターンの予測手法には様々な手法が考案されている。以下にその種類の一部を紹介する。

4.2.1　教師あり機械学習で利用する手法

(1) 線形回帰

教師あり学習において最もポピュラーなのが回帰分析である。その中でも、線形回帰は初歩的な分析手法の一つで、適合性の高い直線を引きながら、最適解を求めていく。説明変数が一つの線形回帰を「単回帰分析」、説明変数が二つ以上の分析を「重回帰分析」と呼び、目的に応じて使い分ける。線形回帰の活用シーンは、相関が高い項目から数値の予測をするような場面などである。例えば、早稲田大学高等研究所の松岡の報告[25)]によると学校水準の学力と社会的経済地位の関連は相関の関係にあるといわれる（図 4.1）。

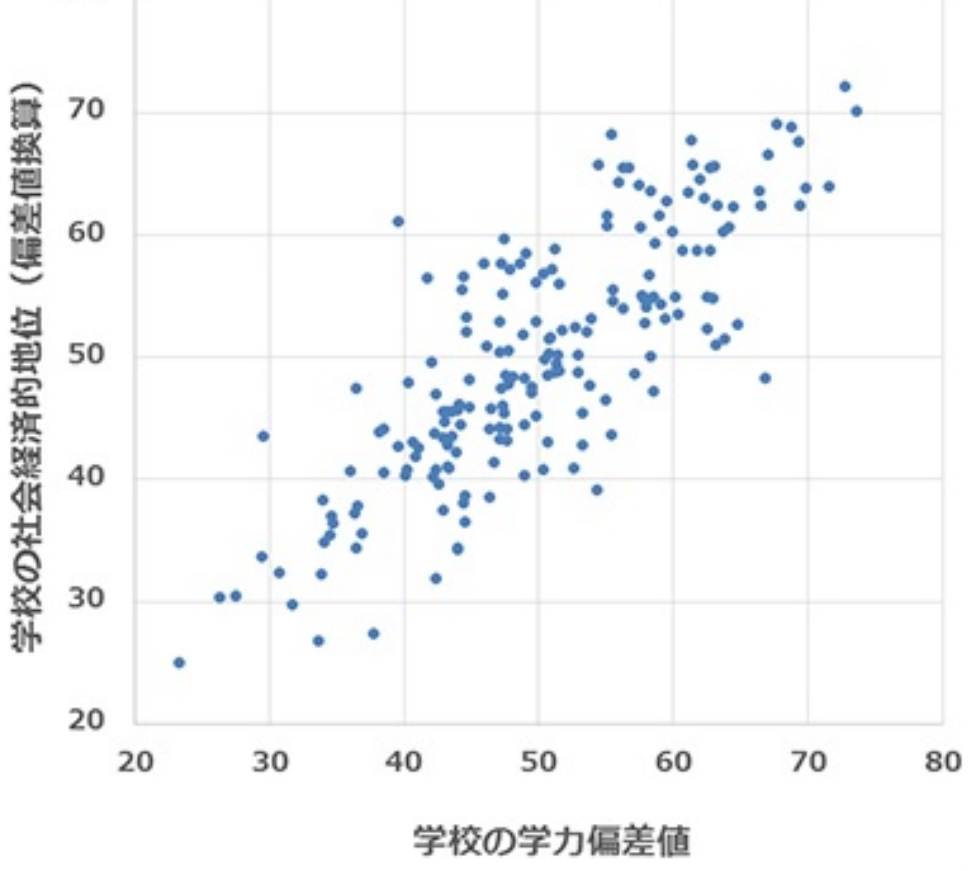


図 4.1　学校水準の学力と社会的経済地位の関連

(2) ロジスティック回帰

ロジスティック回帰は、特定の事象が発生する要因が複数考えられるとき、それぞれの要因から特定の結果がもたらされる確率を計算するための手法である。未だ判明していない結果を予測するのに活躍することはもちろん、すでに判明している結果の確認や説明をする際にも使われる。どのような食事が体重の増減に関連しているのか、飲酒量が病気の罹患率にどう影響しているのかを検証するときなど、活用機会は多岐に渡る。

(3) 決定木

決定木分析は、特定の目的に合わせたツリー構造を形成し、データ分析を進める手法である。特定の行動、例えば新型コロナウイルスのスクリーニング検査を実施するとし

よう。ある集団が過去に集団会食をしたかどうかを問い合わせ、さらに、その後ワクチンを接種済みかどうかを問い合わせて図 4.2 のような結果が得られたとする。

仮に、全体の新型コロナウイルスのスクリーニングによる陽性率が 3 %だとすると、「集団会食をして、ワクチン未接種の場合」に陽性率が 20 %となるので、全体の 6 倍以上にもなる。

回帰分析等に比べて、決定木分析は解析前に必要な前処理が少ないというメリットがある。決定木分析はデータの分布を制限せず、Yes,No のみでデータを分類していくため、欠損値の対応や、標準化や対数変換などの処理が不要である。前処理はデータ分析の仮定において、特に時間のかかる工程の一つであるため、この点において決定木分析は実施しやすい手法であるとも言える。また、決定木は比較的幅広いデータに対してよい性能を発揮でき、汎用性が高いも特徴である。デメリットとしてはあまり分岐を多層にすると過学習が起き予測精度が落ちやすい。そのため階層構造は 4 階層くらいを目安とすることが多く、複雑な予測には向いていない。

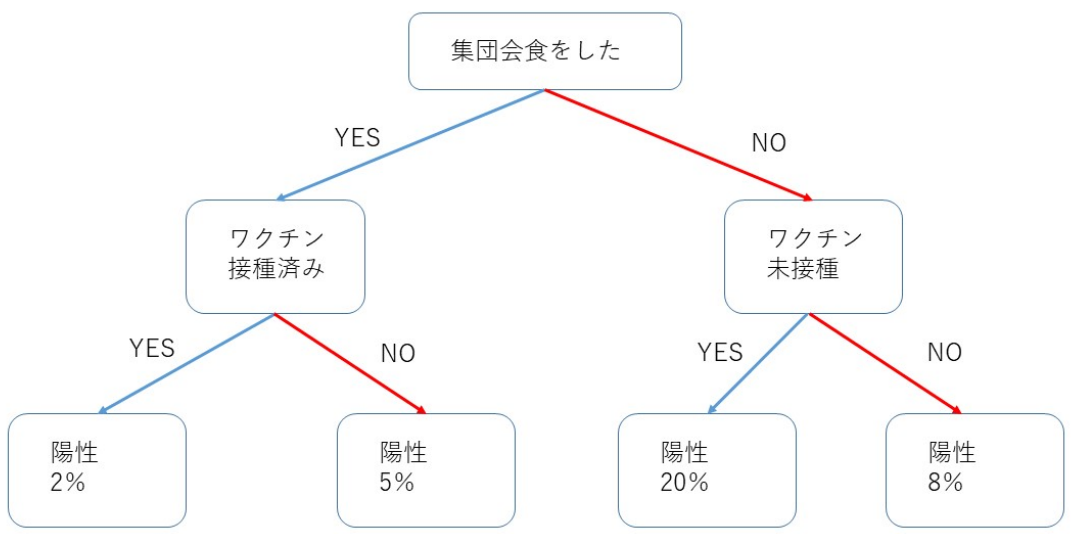


図 4.2　新型コロナウイルスのスクリーニングに決定木分析を活用した例

(4)RandomForest

RandomForest は決定木をたくさん集めて合体させた手法である。決して精度が高いとはいえない弱いモデルをたくさん構築し、これらの予測結果を統合することで高い精度を出す方法である。

(4) サポートベクターマシン（SVM）

サポートベクターマシン [20] は、教師あり学習の代表的なアルゴリズムで、精度の高さやスピードの速さに提唱がある手法である。データを分割する際に現れる直線に最も近い点をサポートベクターと呼び、このサポートベクターを使って直線が上にあるか、下

にあるかを把握し、クラスの分類を実行する。

例えば図 4.3 のような 2 つのグループを分類する線形識別問題があったとすると、サポートベクターマシンでは 2 つを分ける直線を求める。この時、最も良い感じに分けたいときにマージン最大化と呼ばれる、直線と線が最大になる部分を線を引いてクラス分類を行う。また誤識別をコスト関数に追加することでソフトマージンを実現することもできる。

さらに非線形の識別関数も学習可能である。カーネルトリックという手法を用い線形ではない場合でも、次元を増やすことで線形識別問題を解決する（図 4.4)。この場合は非線形サポートベクターマシンと呼ばれる。

サポートベクターマシンは識別能力が高く、少ないパラメータで実行できることから、頻繁に利用されている。

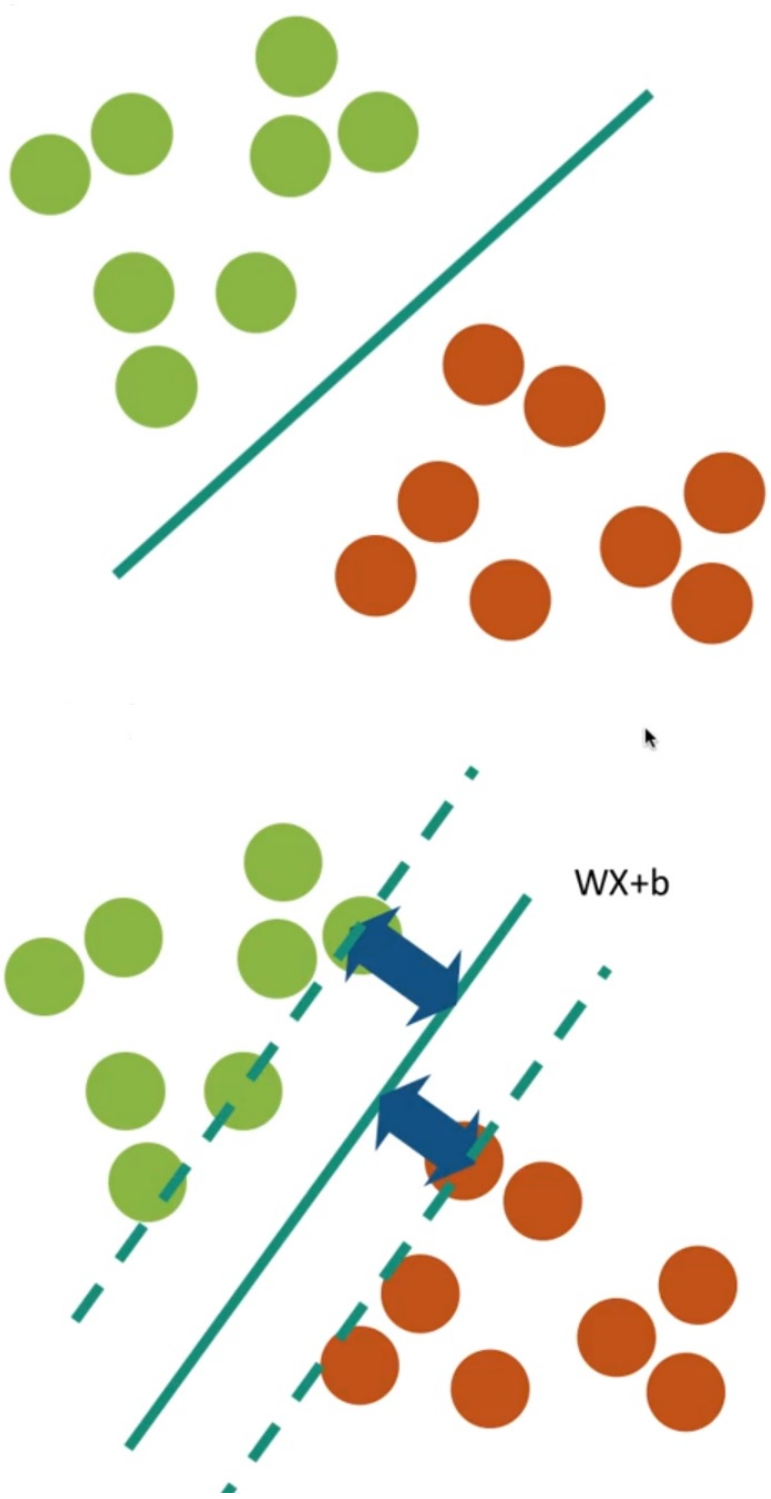


図 4.3　サポートベクターマシン（SVM）で線形識別問題を解く

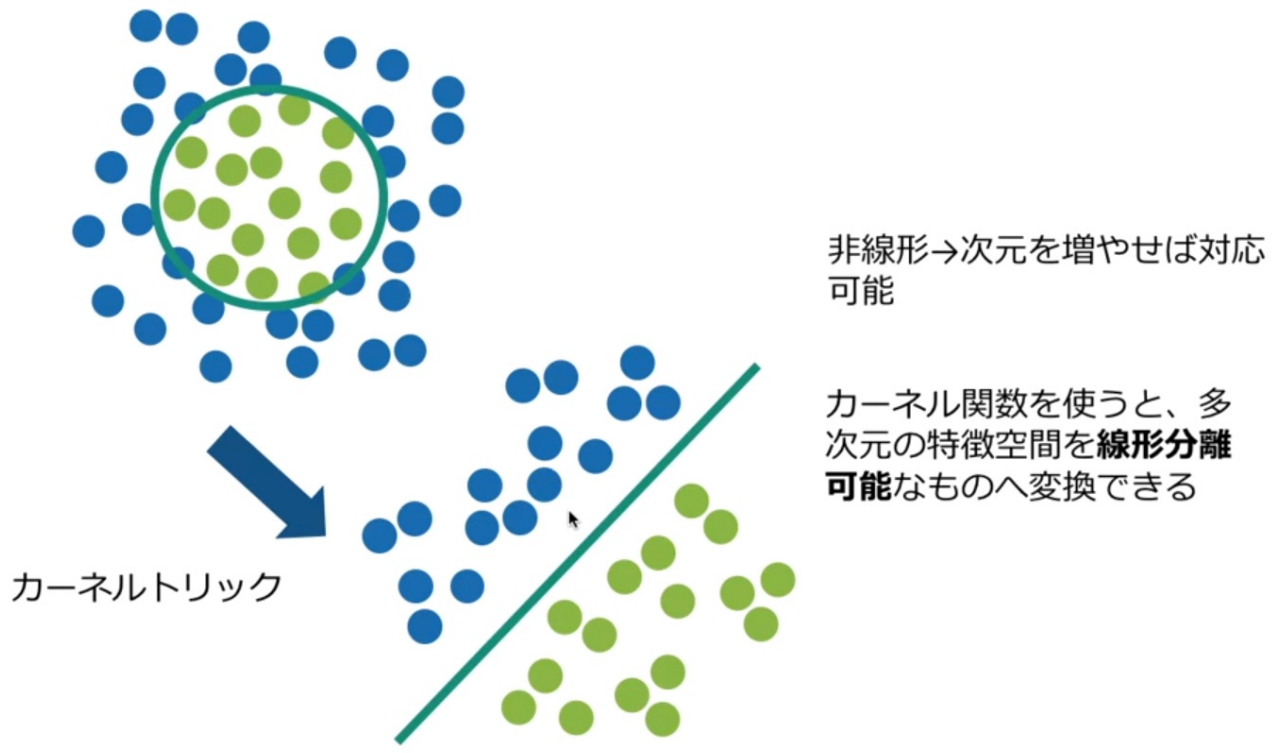


図 4.4　サポートベクターマシン（SVM）で線形識別問題を解く

4.2.2　教師なし機械学習で利用する手法

(1)k 平均法

k 平均法は k-means 法とも呼ばれ、類似する要素同士をクラスタリングする際に用いられる手法である。座標状にさまざまなデータが散らばっている場合、分類の形式や分類するクラスの数を指定してやることで、座標上のデータが条件に合わせて分類される。教師あり学習とは異なり、正解があらかじめ与えられていないため、ユーザーが自ら分析結果を解釈する必要がある。

(1) 主成分分析

主成分分析は、多様な種類のデータを要約し、特徴の把握を促すために用いられる手法である。そのままでは把握が難しい説明変数をより少ない指標に落とし込んでしまうことで、データのわかりやすさを突き詰めることができる。ビッグデータから重要な情報を抜き出す上でも、主成分分析の手法は用いられる。目的に応じて必要な変数を抽出し、機械学習へと応用する。次元を縮小しても重要な要素が削られにくいということから、教師なし学習を実施する上ではポピュラーな手法といえる。

4.2.3　機械学習を用いた問題解決

4.2.4　特徴量抽出

機械学習で問題を解決する際に必要となるのが特徴量抽出である。これは、数値として表現されていない自然言語などのデータを数値に変換する処理である。これらを取得することによってモデルにデータを与えることが可能になる。また、モデルの性能の向

上を目的として数値化したデータに次元削減を行い、データを圧縮するといったアプローチを行うこともある。教師あり学習では人間がこれを与えることが多く、教師なし学習ではモデルがこれを見出すことが多い。

4.2.5　ニューラルネットワーク

　ニューラルネットワークは、ディープラーニングに採用されている分析手法である。人間の脳神経（ニューロン）をモデルにしたネットワーク構造で、多層構造の中にデータをインプットさせることで、特徴量を把握したり、データのパターンを読み込んだりすることが可能である。「入力層」「隠れ層」「出力層」の三層構造で成立しており、隠れ層が増えるほどに精度の高い読み込みが実現する。複数のレイヤーの組み合わせにより構成され、その層の一つ一つは何らかの変換を行う。各層はそれぞれ入力中に対し線形化を行い、例として図 4.5 に示すシグモイド関数の様な活性化関数による出力を返すようになっている。これは人間のニューロンの発火現象を擬似的に再現したものである。昨今機械学習の分野で話題になっているディープラーニングはこのニューラルネットワークの層を多くしたものであるため、日本語では深層学習と言われており、近年は、隠れ層の多層化を実現することで、高度な学習プロセスを実現している。

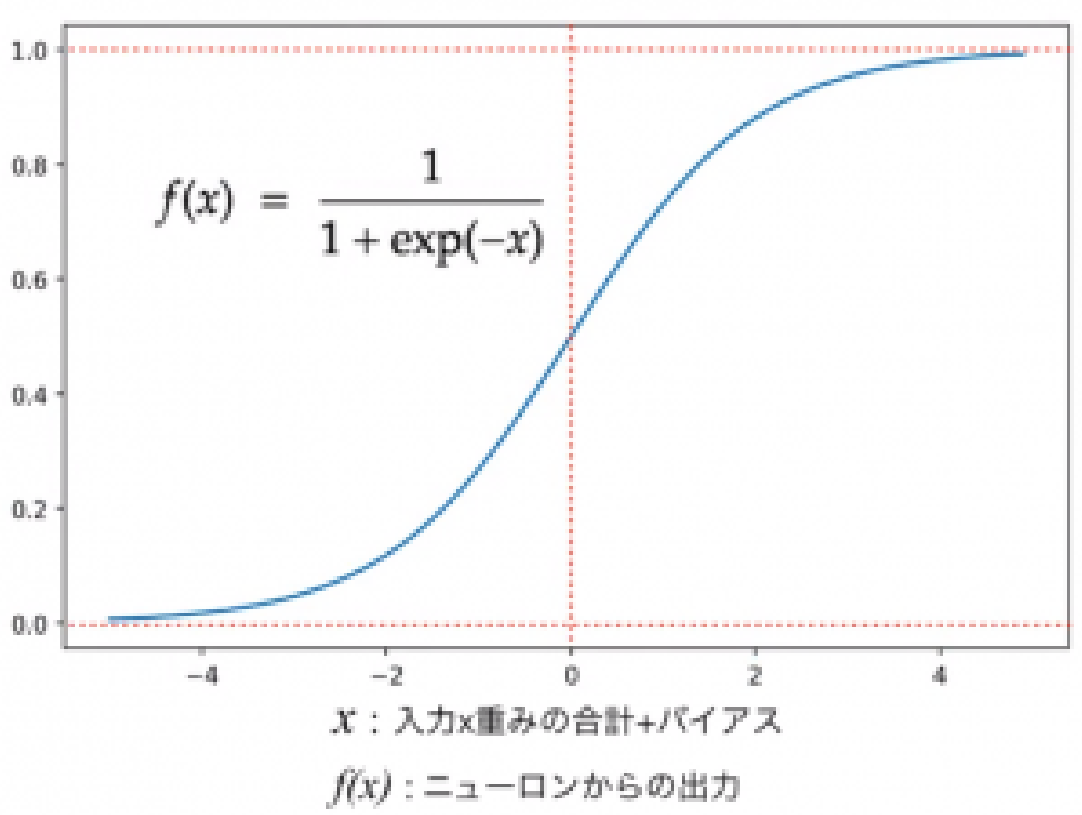


図 4.5　シグモイド関数は、0〜1 の間の値を返す

4.2.6　言語モデル

言語モデルとは、言語を文章の出現しやすさの確率を用いてモデル化したものである。例えば、「本日は頭が痛い」と「本日は爪が痛い」の二つの文章に対して確率を与えるとする。この場合、前者は一般的にもよく使用されるが、後者は稀な場面でしか使用される事のない文であるため、前者の方に後者よりも高い確率を与える。これを行うことにより文章の自然さを確率を用いて表現することができ、文章誤りの訂正などに用いることができる。

4.2.7　オントロジー

オントロジーとはもともとは哲学の用語であり、存在するものの体系的な理論を表していた。人工知能の分野では、存在するものの共通の概念や性質を記述するものとして扱われる。オントロジーの研究は 1990 年代の終わりから 2000 年初めにかけて開始された。知識表現における記述語彙の統一、さらにはその背景に存在する概念体系の明確化が重要視されるようになった。オントロジーの問題解決の対象分野をドメインという。

大江 [21, 22] は、医学オントロジーとその応用システムについて述べた（図 4.6）。将来的な医療オントロジーの活用場面として、テキストマイニングによる知識発見、文章データを含む他施設臨床医学データの疫学的解析（疾患と症状、検査異常の頻度を解析し原因やリスク因子の重要度を解析すること）、類似の臨床経過を辿った患者の検索の他、フリー入力された臨床所見データの表現の揺れを吸収した自動的な ICD10 コーディングや DPC 入力支援に応用できるとした。しいては医療安全確保のための高度な意思決定支援システムや診断支援システムの応用に期待した。一方、実質的な課題を解決していくためには、医療オントロジーが疾患や人体解剖構造だけでなく、日常臨床の場で使用されている医療用語の同義語バリエーションとオントロジーでの語彙との対応も必要と述べた。また、巨大化したオントロジーを進歩し続ける医療知識に合わせてメンテナンスしていく体制の整備や医学文献からの半自動収集が課題であるとした。

臨床医学オントロジーの活用

- 将来の活用例
  - 蓄積された臨床データベースから
    - テキストマイニングによる知識発見
    - 類似症例や類似の疾患経過の検索
    - データ入力時の患者状況に依存した専門用語提示など高度なマンマシンインターフェイスの実現
    - 文章データを含む多施設臨床医学データの疫学的解析
    - フリー入力された臨床所見データの表現の揺れを吸収した自動コーディングや統計処理
    - 意味的相互運用性を維持した施設間のデータ交換やデータ移行
    - 医療安全確保のための高度な意思決定支援システムや診断支援システム

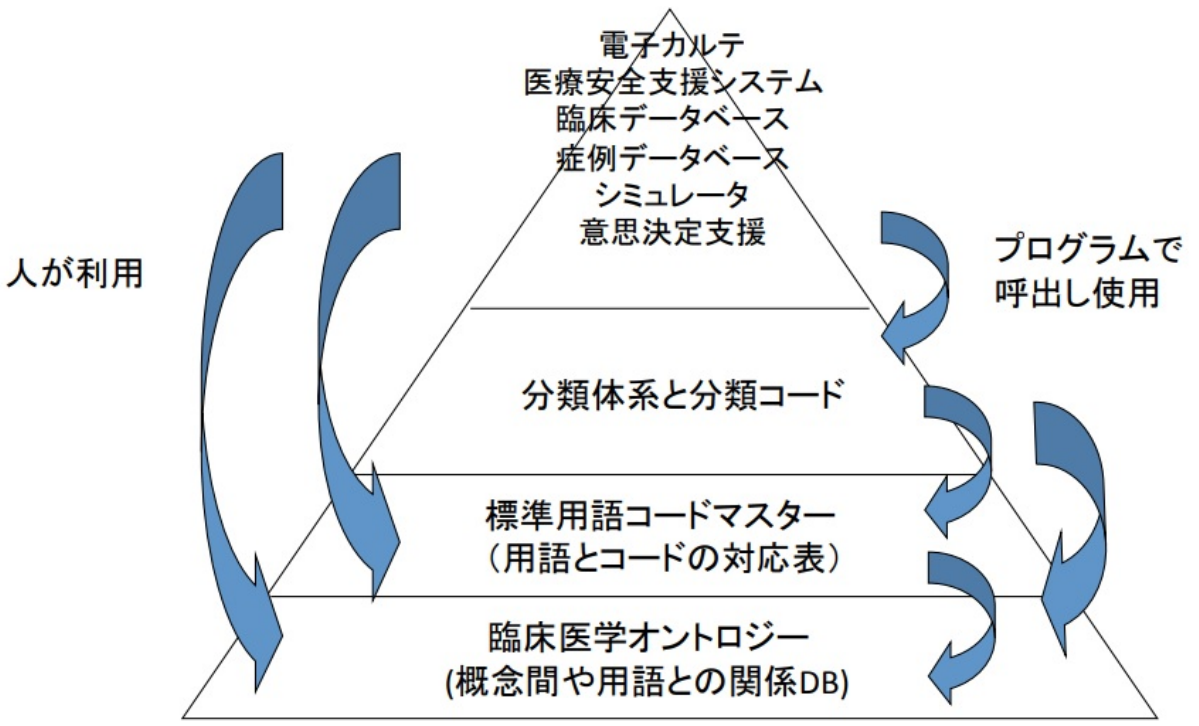


図 4.6　医学オントロジーとその応用システムとの関係 大江和彦 medinfo レポート 2009 より抜粋

第 5 章

自然言語

　日本語に対して機械学習を行う場合自然言語の理解と形態素解析の理解が必要である。この章では形態素解析について記載する。

5.1　自然言語

　自然言語とは日本語や英語、中国語のように人間が日常的に意思疎通を目的として使用する言語である。対の概念としてプログラミング言語などの人工言語が存在する。これら二つの違いは人工言語がコンピュータを対象としているためコンピュータに解釈しやすいような構造になっており厳格に解釈が定められているのに対し、自然言語は「上手（じょうず)」や「上手（かみて)」などの多義語などの解釈に幅のあるものが存在するという点である。

5.2　自然言語の曖昧性

　テキストマイニングにおける問題の原因となっているものの一つとして自然言語の曖昧性が挙げられる。例の一つとして多義語が挙げられ、これは同じ単語であってもそれぞれ持つ意味が異なることを指している。例として「生物（せいぶつ)」と「生物（なまもの)」などが挙げられる。このように単語が二つ以上の意味を持つ別の単語を同じものだと判断してしまうという課題がテキストマイニングに存在する。また、「赤いリンゴの描かれたコップ」というものに対しても「赤い」「リンゴの描かれたコップ」なのか「赤いリンゴの描かれた」「コップ」なのか判別がつきにくいといったように、自然言語の使い方には厳格なルールがないためこのような問題が発生しやすい。

5.3　形態素解析

形態素解析とは自然言語処理を行う上で、処理前に行う品詞に分解する作業である。例えば、「すももももももももものうち」これを分解すると、すもも　も　もも　も　もも　の　うち　になる。このように意味が通るように最小単位に分析する（図 5.1）。

```
すももももももももものうち
すもも　名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も　　助詞,係助詞,*,*,*,*,も,モ,モ
もも　　名詞,一般,*,*,*,*,もも,モモ,モモ
も　　助詞,係助詞,*,*,*,*,も,モ,モ
もも　　名詞,一般,*,*,*,*,もも,モモ,モモ
の　　助詞,連体化,*,*,*,*,の,ノ,ノ
うち　　名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
```

図 5.1　形態素解析ソフト MeCab を利用して形態素解析を行ってみたところ.

英語の場合は、単語毎にスペースが入るので形態素解析の処理は不要である。日本語の場合は、形態素解析ソフトに医療の用語の専門辞書をインポートすることで、意図しない用語の分割を防ぐことができる。医療分野の形態素解析ソフトは MeCab[26] が広く採用されており、専門用語の辞書も用意されていることから、本研究でも MeCab を利用することとした。

5.3.1　MeCab

MeCab は工藤拓が開発した形態素解析ソフトである。専門用語を扱う場合は MeCab 用に別途辞書を用意することで意図しない単語の分解を防ぐことができる。カルテの経過記録などの医療文書を自然言語解析する場合にメリットが大きい。

Mecab 専用のユーザー辞書を用意することで、品詞の分解を防ぐことができる。辞書フォーマットは、「表層形, 左文脈 ID, 右文脈 ID, コスト, 品詞, 品詞細分類 1, 品詞細分類 2, 品詞細分類 3, 活用型, 活用形, 原形, 読み, 発音」と、カンマで区切られた csv ファイルで成り立つ。

5.3.2　MeCab 用医療辞書

MeCab 用の医療辞書としてオープンに提供されているものには「万病辞書」[27] と ComeJisyo[28] がある。その他オリジナルでユーザ辞書を作成することも可能である。

6.1.2.1 万病辞書

日本語形態素解析用の電子医療用語辞書である。協力医療機関で得られたテキスト情報から、約 160 万以上の症状・病名に関する語を抽出したもの。そのうち、特に頻出する約 36.3 万の症状・病名に関する語を抽出し、既存辞書である ICD10 対応標準病名マスターに含まれている語と紐づけしたものを収載している。ICD10 対応標準病名との対応付けとして、症状・病名に関する語に対して、その語に最も近い ICD10 コードならびに ICD10 対応標準病名が付与されている。

6.1.2.2 ComeJisyo

日本語形態素解析用の電子医療用語辞書である。主に看護用語を中心にカバーしているが、看護だけにとどまらず一般の医療領域全般の用語も含む。現在、看護文書に含まれる 5 万語、看護学教科書の索引より抽出した 4 万語、看護師国家試験から抽出した 1 万語、ウェブで公開される用語辞書から抽出した 3 万語について、品詞、読み仮名、形態素解析のための連接コストが付与されている。

6.1.2.2 オリジナルのユーザー辞書 (ManbyoCome5)

本研究では万病辞書と ComeJisyo の 2 種類の辞書を 1 つのファイルにまとめ、オリジナルのユーザー辞書として MeCab から利用できるようにした。原形に標準病名を登録した（図 5.2)。形態素解析時に原形を出力することで表記揺れの問題を改善できる。

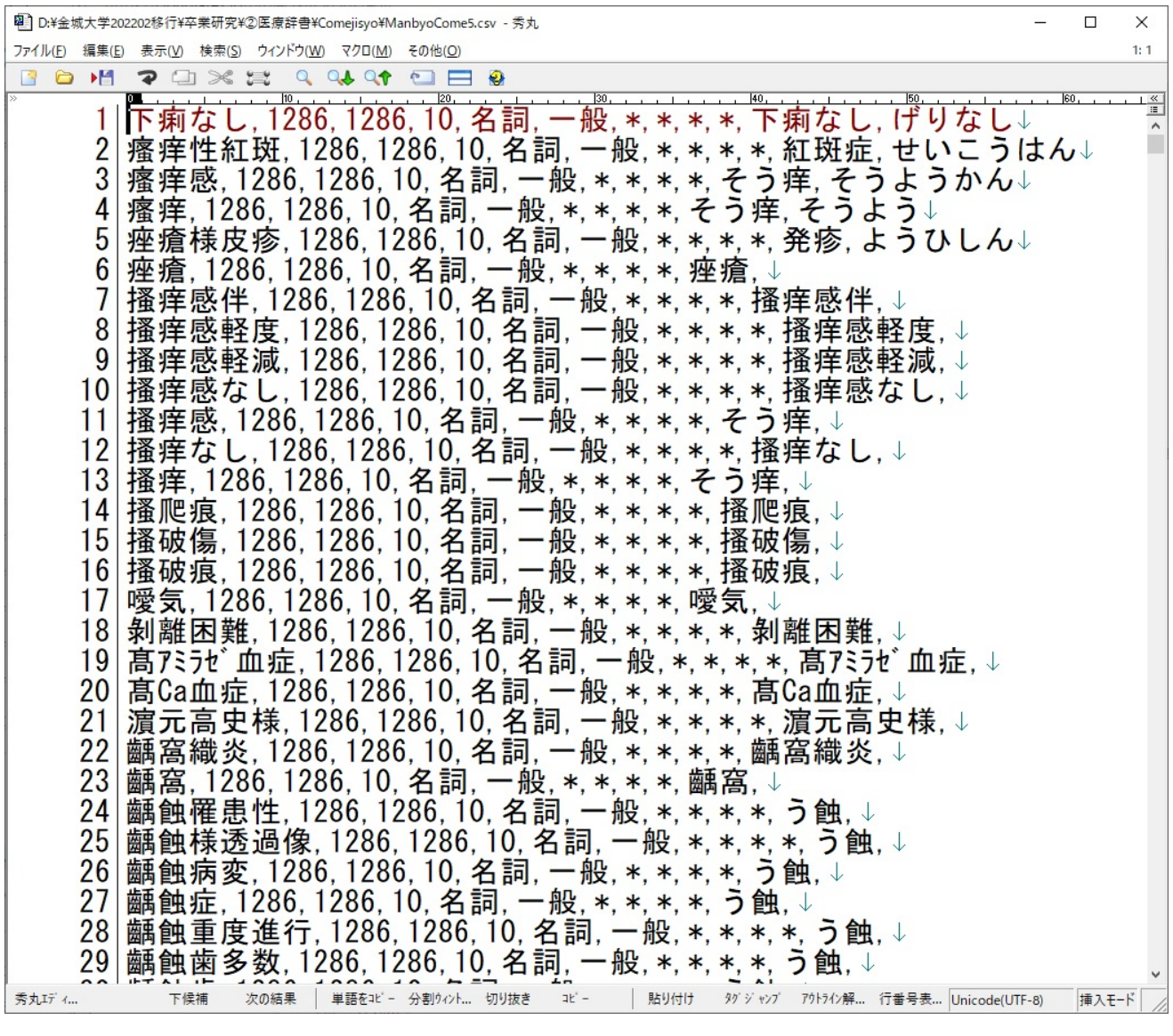


図 5.2　万病辞書と Comejisyo を融合させたオリジナルのユーザ辞書 (ManbyoCome5) の内容

第 6 章

ベクトルと行列

自然言語処理では、単語をベクトルに変換する。この章では、ベクトルに関する数学の知識 [23] を記載する。

6.1 ベクトル

6.1.1 ベクトル

ベクトルは平面上で向きを表した量で

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

などとかける。この場合 x 方向に 1、y 方向に 2 という意味となる。なお、同様の意味で

$$(1, 2)$$

と横に表記する方法もあるが、本論文では縦に並べる表記とする。ベクトル自体は 2 次元に限る必要はなく何次元でも良い。二次元であれば成分が 2 個、三次元であれば 3 個というように次元の数だけ成分を持つ。ベクトルを文字で表記する際は、数字と区別するために $\mathbf{a}$ または $\vec{a}$ と表す。機械学習では $\mathbf{a}$ が用いられることが多い。以下のような表記がよくなされる。

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$$

すなわち i 成分目を a_i のように記載する。

6.1.2 ベクトルの大きさ

ベクトルの大きさは原点から各成分の距離に対応していて、$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$ の場合以下のように定義される。

$$\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} = \sqrt{\sum_{i=1}^{n} a_i^2}$$

ベクトルの長さは $|\mathbf{a}|$、$\|\mathbf{a}\|$、またまた単に a と表すなど様々な書き方が存在する。なお、この距離の定義をユークリッド距離と呼ぶ。

6.1.3　ベクトルの定数倍

ベクトル $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$ の定数倍を実数 k を用いて

$$k\mathbf{a} = \begin{pmatrix} ka_1 \\ ka_2 \\ \cdots \\ ka_n \end{pmatrix}$$

で定義する。このときこのベクトルの長さは

$$\|k\mathbf{a}\| = \sqrt{\sum_{i=1}^{n} k^2 a_i^2} = |k| \sqrt{\sum_{i=1}^{n} a_i^2} = |k| \, \|\mathbf{a}\|$$

のように k 倍される。ベクトルが k 倍されたとき、ベクトルの長さが k 倍される。$k < 0$ のときは、ベクトルの向きが逆になる。

また、$k = \frac{1}{\|\mathbf{a}\|}$ とすると、長さ 1 のベクトルになる。長さ 1 のベクトルのことを単位ベクトルと呼び、ベクトルの長さを 1 にする行為を正規化と呼ぶことがある。また全ての成分が 0 のベクトルをゼロベクトルと呼び、任意のベクトルで $k = 0$ とした場合に対応する。

$$\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}$$

6.1.4　ベクトルの加減

ベクトル $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$ と $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$ の加減は各成分どうしの加減で定義される。

$$\mathbf{a} \pm \mathbf{b} = \begin{pmatrix} a_1 \pm b_1 \\ a_2 \pm b_2 \\ \cdots \\ a_n \pm b_n \end{pmatrix}$$

6.1.5　内積

ベクトル $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$ と $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$ の内積と呼ばれる量は、各成分をかけたものを全て足した数で定義され、$\mathbf{a} \cdot \mathbf{b}$、$\langle \mathbf{a}, \mathbf{b} \rangle$、$(\mathbf{a}, \mathbf{b})$ のように書き、以下のように定義される。

$$(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} a_i b_i$$

6.2　行列

6.2.1　行列

数を長方形状に並べたものを行列という。例えば以下のようなものをいう。

$$\begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$\begin{pmatrix} 2 & 4 & \frac{1}{2} \\ 0.2 & 1 & 1 \end{pmatrix}$$

行列では縦並びを行、横並びを列という。上記 1 つ目の行列は 2 行 1 列、2 つめは 2 行 3 列の行列と表現され (2×1)、(2×3) などと表現される。$(m \times n)$ の行列は便宜上以下のように表されることが多い。

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{m1} & a_{n2} & \cdots & a_{mn} \end{pmatrix}$$

すなわち $(m \times n)$ の i 行 j 列目の成分を a_{ij} と表す。また、行列はこの場合 A などと大文字のアルファベットで置かれることが多い。またベクトルと同様に数と区別するために $\mathbf{A}$ のように記載することもある。また、成分がすべて 0 の行列をゼロ行列と呼び O と書かれる。

$$O = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

行数と列数が等しい行列を正方行列という。$(n \times n)$ の例を示す。

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

正方行列で、対角成分 (行数と列数が等しい成分) が 1、残りが 0 の行列を単位行列と呼ぶ。単位行列は I で書かれる。

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

行列の行と列を入れ替えた行列を転置行列と呼び A の転置行列を、A^T、${}^tA$、A^t など様々な形で表されるが、機械学習の分野では A^T と書かれることが多い。

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{m1} & a_{n2} & \cdots & a_{mn} \end{pmatrix}$$

の転置行列は

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1m} \\ a_{12} & a_{22} & \cdots & a_{2m} \\ \cdots & & & \\ a_{1m} & a_{m2} & \cdots & a_{nm} \end{pmatrix}$$

となる。例えば

$$\begin{pmatrix} 2 & 4 & \frac{1}{2} \\ 0.2 & 1 & 1 \end{pmatrix}$$

の転置行列は

$$\begin{pmatrix} 2 & 4 & \frac{1}{2} \\ 0.2 & 1 & 1 \end{pmatrix}^T = \begin{pmatrix} 2 & 0.2 \\ 4 & 1 \\ \frac{1}{2} & 1 \end{pmatrix}$$

となる。

6.3 行列の演算

6.3.1 加減

行列の加減は同じ行列数の行列 A、B で定義され、

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{m1} & a_{n2} & \cdots & a_{mn} \end{pmatrix}$$

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & & & \\ b_{m1} & b_{n2} & \cdots & b_{mn} \end{pmatrix}$$

に対して、同じ成分どうしの加減で定義される。

$$A \pm B = \begin{pmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} & \cdots & a_{1n} \pm b_{1n} \\ a_{21} \pm b_{21} & a_{22} \pm b_{22} & \cdots & a_{2n} \pm b_{2n} \\ \cdots & & & \\ a_{m1} \pm b_{11} & a_{m2} \pm b_{m2} & \cdots & a_{mn} \pm b_{mn} \end{pmatrix}$$

例を示す。

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 & 0 \\ -1 & -2 & -3 \end{pmatrix}$$

$$A + B = \begin{pmatrix} 2+2 & 3+1 & 4+0 \\ 5-1 & 6-2 & 7-3 \end{pmatrix} = \begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & 4 \end{pmatrix}$$

$$A - B = \begin{pmatrix} 2-2 & 3-1 & 4-0 \\ 5-(-1) & 6-(-2) & 7-(-3) \end{pmatrix} = \begin{pmatrix} 0 & 2 & 4 \\ 6 & 8 & 10 \end{pmatrix}$$

6.3.2　乗法

次に行列の積 $C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & & & \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{pmatrix}$ と $D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2p} \\ \cdots & & & \\ d_{n1} & d_{n2} & \cdots & d_{np} \end{pmatrix}$ の積について考える。行列の積 CD は C の行と列を取り出して取り出した行と成分で内積を取ればよい。

$$CD = \begin{pmatrix} \cdots & & & \\ c_{i1} & c_{i2} & \cdots & c_{in} \\ \cdots & & & \\ \cdots & & & \end{pmatrix} \begin{pmatrix} \cdots & \cdots & d_{1j} & \cdots \\ \cdots & \cdots & d_{2j} & \cdots \\ \cdots & & & \\ \cdots & \cdots & d_{nj} & \cdots \end{pmatrix}$$

$$= \begin{pmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & c_{i1}d_{1j} + c_{i2}d_{2j} + \cdots + c_{in}d_{nj} & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

行列 C は $(m \times n)$ の行列、行列 D は $(n \times p)$ の行列であるが、その積は $(m \times p)$ となる。C の行数、D の列数が積の計算結果となる。

$$(m \times n)\ \ (n \times p) \to (m \times p)$$

例を示す。

行列 $\begin{pmatrix} 2 & 4 & 1 \\ -1 & 1 & 1 \end{pmatrix}$ と $\begin{pmatrix} 0 & 1 \\ -1 & 2 \\ 1 & -3 \end{pmatrix}$ の積は次のようになる。

$$\begin{pmatrix} 2 & 4 & 1 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 2 \\ 1 & -3 \end{pmatrix}$$

$$= \begin{pmatrix} 2\times 0 + 4\times(-1) + 1\times 1 & 2\times 1 + 4\times 2 + 1\times(-3) \\ (-1)\times 0 + 1\times(-1) + 1\times 1 & (-1)\times 1 + 1\times 2 + 1\times(-3) \end{pmatrix} = \begin{pmatrix} -3 & 7 \\ 0 & -2 \end{pmatrix}$$

と計算できる。

6.3.3　内積の表現方法

この積の定義に着目すれば、ベクトル $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$ と $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$

の内積は次のようにかける。

$$(\mathbf{a}, \mathbf{b}) = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix} = \mathbf{a}^T \mathbf{b}$$

とかける。

6.3.4　逆行列

行列 A にかけた時に単位行列 I となる行列を逆行列と呼び A^{-1} と表す。

$$A^{-1}A = AA^{-1} = I$$

行列 $A = \begin{pmatrix} 1 & 1 \\ 3 & 4 \end{pmatrix}$ の逆行列は $A^{-1} = \begin{pmatrix} 4 & -1 \\ -3 & 1 \end{pmatrix}$ である。

$$A^{-1}A = \begin{pmatrix} 4 & -1 \\ -3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$AA^{-1} = \begin{pmatrix} 1 & 1 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 & -1 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

第 7 章

単語のベクトル化

この章では実際に単語をベクトル化する際の関連項目について記載する。コンピュータに自然言語を処理する際には、単語をどの様に表現するのかを考える必要がある。一般的に、自然言語処理では単語を数値表現に変換して処理する。理由は、自然言語処理で利用される機械学習手法では、単語を文字列としてそのまま扱うのが難しいからである。

近年は、単語の分散表現が利用されることが多くなってきた [24)]。単語の分散表現とは、１つの単語を数百次元程度のベクトルで表す方法である。単語の意味を捉えるような性質をもつ。テキストから単語に分割するには、形態素解析を行う。そして分割した単語に対しベクトル化を行う。

7.1 N-gram ベクトル

N-gram ベクトルはテキストを n-gram にて表す手法である。ここで n-gram とは連続する n 個のトークン (単語や文字など) の事である。例えば「the cat is out of the bag」という文字を n-gram で表現すると、n=1 の場合は、['the','cat','is','out','of','the','bag']、n=2 の場合は ['the cat','cat is','is out','out of','of the','the bag'] のようにテキストを表現できる。特に n=1 の場合をユニグラム (uni-gram)、n=2 の場合をバイグラム (bi-gram) と呼ぶ。

テキストを n-gram に分割したら、機械学習アルゴリズムが理解できるように数値のベクトルに変換する必要がある。そのためにはまず、各 n-gram に対して重複のないように数値を割り当てる。これを語彙 (vocabulary) と呼ぶ。例えば、先の文をユニグラムにしたものについて以下の様な語彙を作ることができる。

語彙の例

text:'the cat is out of the bag'

Vocabulary:'the':0,'cat':1,'is':2,'out':3,'of':4,'bag':5

語彙を作成したら、テキストを Bag-of-Ngrams(BoW) と呼ばれる形式でベクトル表現

にする。BoW の考え方はシンプルで、テキストに n-gram が含まれているかどうかだけを考慮し、その並び方は考慮しない。特にユニグラムのものを BoW という。BoW で表すベクトルにはいくつかのバリエーションがある。ここでは n=1 の場合を対象に、以下のバリエーションについて考える。

・One-hot エンコーディング
・Count エンコーディング
・tf-idf

7.2　One-hot エンコーディング

単語の One-hot エンコーディングとは単語をベクトルで表現する方法の一つである。one-hot エンコーディングではある要素のみが「1」でその他の要素が「0」であるようなベクトルで単語を表現する。各次元に「1」か「0」を設定することで「その単語か否か」を表す。次元数はボキャブラリ数と等しくなる。ここでボキャブラリとは単語の集合を表す。

例えば「the cat is out of the bags」という文を BoW で表現するとする。ここで、語彙は全部で 8 単語で {'are':0,'bag':1,'cat':2,'dogs':3,'is':4,'of':5,'out':6,'the':7} のような割り当てになっていたとする。そうすると、文書は図 7.1 のように 8 次元のベクトルで表現することができる。

0	1	1	0	1	1	1	1
are	bag	cat	dogs	is	of	out	the

図 7.1　One-hot エンコーディング

One-hot エンコーディングの欠点は、ベクトル間の演算で意味のある結果を得られない点である。例えば、単語間の類似度を計算するために内積を取るとする。one-hot エンコーディングとはその性質上、異なる単語間の内積を取った結果は「0」になる。しかし、単語には「犬」と「猫」は似ているが、「猫」と「石」は似ていないといった関係があるはずである。このような関係を One-hot エンコーディングでは扱うことができない。

さらに、1 単語に 1 次元を割り当てるため、新しい単語を追加するたびに、ベクトルの次元を増やさなければならない。次元が増えると、元の次元数のベクトルを与えて学習させたモデルを学習しなおす必要が生じる点も欠点といえる。

では実際に One-hot エンコーディングでの Bow を Python を用いて実装してみる。今

回は、scikit-learn に実装されている CountVectorizer を利用した。

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(binary=True)
docs = ['the cat is out of the bag','docs are']
bow = vectorizer.fit_transform(docs)
bow.toarray()
```

One-hot エンコーディングで得られたベクトル値を表 7.1 に示す。

表 7.1　One-hot エンコーディングで重みづけした場合のベクトル値の出力結果

are	cat	docs	is	of	out	the
0	1	1	0	1	1	1
1	0	0	1	0	0	0

the cat is out of the bag の文が、0, 1, 1, 0, 1, 1, 1, 1 に変換された。ただし、このままでは単語とインデックスの対応が分からない。対応のためには CountVectorizer の属性である vocabulary_を参照することで得られる。

```
vectorizer.vocabulary_
{'the': 7, 'cat': 2, 'is': 4, 'out': 6, 'of': 5, 'bag': 1, 'docs': 3, 'are': 0}
```

fit_transform メソッドの中では、まず vocabulary_に単語とインデックスの対応を作成した後、作成した vocabulary_を用いて BoW 表現に変換している。

7.3　Count エンコーディング

Count エンコーディングでは、ある単語がテキストに存在するか否かだけでなく、その頻度を考慮してベクトルを作成する。

つまり、頻度が高い単語を重視するようなベクトルができる。例えば、先ほどと同じ語彙で同じ文をベクトル化すると、以下の図 7.2 のようになる。

0	1	1	0	1	1	1	2
are	bag	cat	dogs	is	of	out	the

図 7.2　Count エンコーディング

Count エンコーディングでの BoW を Python で実装してみる。One-hot エンコーディングとの違いは、CountVectorizer で binary パラメータに False を渡している点だけである。

```
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(binary=False)
docs = ['the cat is out of the bag','docs are']
bow = vectorizer.fit_transform(docs)
bow.toarray()
```

表 7.2　Count エンコーディングで重みづけした場合のベクトル値の出力結果

are	cat	docs	is	of	out	the
0	1	1	0	1	1	2
1	0	0	1	0	0	0

「the」の値が頻度を考慮して 2 となっている。

7.4　ti-idf

Count エンコーディングの手法の欠点としては、単語の出現回数のみによって単語に重みづけをしていることである。出現回数のみによって重みづけすると、「the」のようにどのテキストでもよく出現する単語に大きな重みが割り当てられてしまう。

tf-idf によるベクトル化は Count エンコーディングの欠点を軽減する手法である。tf-idf では、単語の出現頻度 tf(Term Frequency：単語の出現頻度) をそのまま使うのではなく、ある単語が出現する文書数の逆数 idf（Inverse Document Frequency：逆文書頻度）をかけて単語の重みを表現する。式にすると以下の式となる。

$$tf - idf(t, d) = tf(t, d) \times idf(t, d) \tag{7.4.1}$$

ここで、$tf(t,d)$ は文書 d における単語 t の出現頻度を表している。一方 $idf(t,d)$ は以下のように計算される。ここでは $df(t,d)$ は単語 t が出現する文書数、N は全文書数である。

$$idf(t,d) = log \cdot \frac{N}{1+df(t)} \tag{7.4.2}$$

要するに tf-idf が何を表現しているかというと、頻繁に出現する単語を重要とみなしつつ、多くの文書に出現する単語は重要ではないということである。例えば、ある文書内で「the」の出現頻度が高いと考えられるが、同時に多くの文書に出現すると考えれる。そのためその分を割り引いて重み付けを行う。図 7.3 は tf-idf を計算した結果である。先の Count エンコーディングとくらべて「the」の重みづけが相対的に小さくなっていることが確認できる。

0	0.38	0.38	0	0.38	0.38	0.38	0.54
are	bag	cat	dogs	is	of	out	the

図 7.3　tf-idf

次に Python を用いた tf-idf を実装例を示す。scikit-learn に実装されている TfidVectorizer を利用した。

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer
docs = ['the cat is out of the bag','the docs are']
tfidf = vectorizer.fit_transform(docs)
```

次に if-idf で重みづけたベクトルを求めた。出力結果は表示 7.1 のとおりである。Count エンコーディングと比較して、「the」の重みが相対的に下がっていることが確認できる。

```
import pandas as pd
vocab = vectorizer.get_feature_names()
pd.DataFrame(tfidf.toarray(), columns=vocab).round(2)
```

表 7.3　tf-idf で重みづけした場合のベクトル値の出力結果

	are	cat	docs	is	of	out	the
0	0.00	0.38	0.38	0.00	0.38	0.38	0.54
1	0.63	0.00	0.00	0.63	0.00	0.00	0.45

7.5　単語の分散表現

1950 年代に提唱された、「単語の意味はその単語が使われた周囲の文脈によって決まる」という分布仮説に基づいて、大量の文書情報から単語をベクトルで表現する手法を単語の分散表現という。

局所表現が、各概念毎に一対一で表現するのに対し、分散表現は、人間が新しいことを記憶する際に、既に知っていることと関連させて記憶するという脳神経科学の知見を応用し生まれた。一つの単語は 200 次元〜1000 次元程度のベクトルで表現され、分散表現では文章中に出現する単語数が膨大であってもデータサイズを抑えられ、One-hot 表現よりも計算コストが抑えることができる。

例えば、(PYTHON,BASIC,COBOL) の 3 単語を分散表現で表すと図 7.4 のようになる。

PYTHON	0.52	0.21	0.37	・・・	0.01
BASIC	0.47	0.23	0.33	・・・	0.04
COBOL	0.49	0.01	0.45	・・・	0.12

図 7.4　分散表現のイメージ、ここで図中の値は適当な値を設定している.

また単語をベクトル化し表現することで、単語間の類似度を計算できるようになった (図 7.5)。分散表現により、ある概念と他の概念との類似性を紐づけながら、ベクトル空間上に表現できる。

類似度を計算できるだけでなく、ベクトルを加算することで、単語の意味を捉えるかのような演算を行うことができる。筆者らは先行研究 [29] にて、図 7.6 は腎臓疾患症状である蛋白尿と処方薬のプレドニンを加算して、単語の類似度を演算し、関連病名の導出を試みたところ、IgA 腎症とネフローゼ症候群が近いベクトル値として導出された。

PYTHON	0.71	0.12	0.13	0.25	0.63
	×	×	×	×	× → ∑ = 0.99
RUBY	0.66	0.15	0.04	0.27	0.68
	×	×	×	×	× → ∑ = 0.62
ORANGE	0.36	0.60	0.43	0.55	0.18

図 7.5　単語の分散表現することで単語間の類似度が求められる.

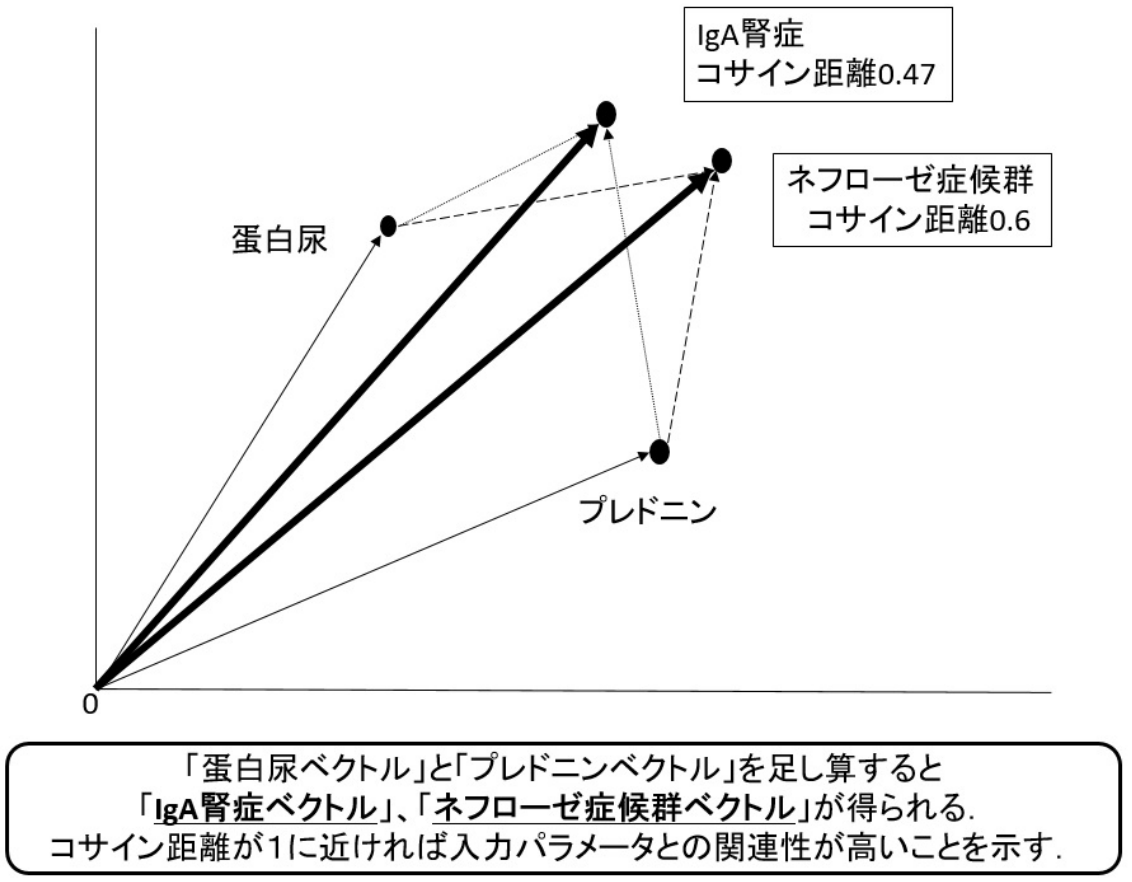


図 7.6　単語ベクトルの演算によって、症状と処方薬の加算から関連病名を抽出できる可能性がある。

次元数は固定なのでボキャブラリ数が増えても各単語の次元数は増やさずに済む。分散表現では 200 次元〜1000 次元程度のベクトルで表現され、この次元数はボキャブラリ数とは関係なく固定できる。従って新しい単語を追加しても次元数を増やさなくても良く、計算コストの増大を抑えれるというメリットがある。

分散表現が重要な理由として、用いる分散表現によってタスクの性能が左右される点が挙げられる。意味表現学習ではニューラルネットワークベースの手法を用いている。ニューラルネットワークへ入力するのは単語分散表現が使われる。単語の意味をよりよく捉えている意味表現ベクトルを入力パラメータとすることで、タスク性能が向上が実現している。

7.6　Word2Vec

Word2Vec は 2013 年に google の研究者、Tomas Mikolov らによって発表された分布仮設に則り大量の文書情報から単語のベクトル表現の学習方式である [30, 32]。2 層のシンプ

ルなニューラルネットワークで構成され、分散表現学習に大容量のメモリを必要とせず、高速に処理できることが特徴である。

Word2Vec の基本的な学習の仕組みは、まず文書情報からを単語列として $w_1, w_2, \cdots, w_T$ を求める。T は文書情報に含まれる単語数である。次に、ある位置 t で出現する単語 w_t に対し、前後 δ 個の単語列を文脈 $C_{wt} = (W_{t-\delta}, \cdots, W_{t-1}, W_{t+1}, \cdots W_{t+\delta})$ とする。文脈 C_{wt} から単語 W_t を予測する条件付き確率分布関数 $P_\Theta(W_t|C_{wt})$ を定義し、式 (7.1) にある対数尤度関数 L が最大となるように、$P_\Theta(W_t|C_{wt})$ を学習する。

$$L = \sum_{t=1}^{r} logP_\Theta(W_t|C_{wt}) \tag{7.1}$$

skip-gram（Continuous Skip-Gram Model）と CBOW（Continuous Bag-of-Words Model）の 2 種類のモデルが用意されており、Word2Vec とはこの 2 つのモデルの総称のことである。

7.6.1 skip-gram

skip-gram 法は、中心のある単語から周辺の単語を予測する手法である。skip-gram 法で行われる学習は教師あり学習である。入力として中心語を与え、その周辺語の予測を出力する。この学習を通じて、ネットワークにある単語の周囲に、どのような単語が現れる可能性が高いのかを学習させる。文脈 C_{wt} に含まれる語は互いに独立であると仮定して、$logP_\Theta(W_t|C_{wt}$ を、文脈単語 c から対象とする単語 W_t を予測する条件付き確率分布関数 $P_\Theta(W_t|C_{wt})$ の積に分解する。よって式 (7.1) は次式に変形できる。

$$L = \sum_{t=1}^{r} \sum_{c \in c_{wt}} logP_\Theta(W_t|C) \tag{7.2}$$

そして、条件付き確率分布関数を次のように対数双線形モデルを用いて定式化する。

$$P_\Theta(W_t|C_{wt}) = \frac{exp(v_c \cdot \tilde{v}'_w)}{\sum_{w' \in v} exp(v_c \cdot \tilde{v}_{w'})} \tag{7.3}$$

ここで、v_c は文脈内にある単語 c のベクトル表現、$\tilde{v}'_w$ は予測単語 w のベクトル表現、そして V はコーパス全体の語彙集合である。

式 (7.3) の右辺の分母には、全語彙集合 V に対する内積と指数関数の計算が現れるため、大規模なコーパスに対して膨大な計算量となる。そこで $|V|$ に比例しない計算量でこの計算を近似する様々な手法が考案されている [31)]。なかでも skip-gram は、負例サンプリングという手法を用いて大幅に計算量を削減している。

負例サンプリングでは、学習データに現れる単語・文脈ペア $\langle w_t, c \rangle$ ごとにランダムに K 個の擬似負例単語 $\langle \tilde{w}_t, c \rangle$ を生成し，それらを識別するように学習する。具体的には，

正例$\langle w_t, c\rangle$に対しては 1，負例$\langle \tilde{w}_t, c\rangle$に対しては 0 を予測するロジスティック回帰モデルで近似する。

7.6.2　CBOW

CBOW も skip-gram と同様、教師あり学習である。skip-gram 法が中心語から周辺語を予測するのに対し、CBOW は周辺の単語から中心語を予測する。この場合の入力は周辺語、出力は中心語となる (図： Word2Vec のニューラル言語モデル)。

CBoW モデルでは，式 (6.2) のように文脈 C_wt に対する条件付確率分布関数を文脈語 c に対する条件付確率分布関数の積に分解せず，文脈内にある単語 c のベクトル表現 v_c の和 $v_{ct} = \sum_{c \in c_t} v_c$ を用いて次式のように条件付き確率分布関数 $P_\Theta(W_t|C_{wt})$ を定式化する。

$$P_\Theta(W_t|C_{wt}) = \frac{exp(v_c \cdot \tilde{v}_{w_t})}{\sum_{w' \in V} exp(v_{c_t} \cdot \tilde{v}_{w'})} \tag{6.4}$$

これ以降の計算は、skip-gram と同様に行う。

7.6.3　ニューラルネットワークと単語ベクトルとの関係

式 6.3 で唐突に P_Θ が出てきたが、これはコーパスを使って最尤推定したいパラメータ Θ である。そして、これらの単語のベクトル表現は入力層、隠れ層、出力層からなるニューラルネットワークの重み行列になっている。図は、Word2Vec のニューラルネットワークを概念的に描いた図である。この図において、$|V|$ はコーパス全体の語彙数で、N は隠れ層のニューロン数を表している ($N << |V|$)。また、$W_{|V| \times N}$ は入力層と隠れ層の間の重み行列、$W'_{N \times |V|}$ は隠れ層と出力層の間の重み行列である。入力層から文脈単語 c の One-hot ベクトルが入力され、重み行列 $W_{|V| \times N}$ を使って隠れ層（中間層）への入力値が計算される。前述の如く、One-hot ベクトルとは、(0,1) を要素とする $|V|$ 次元ベクトルで、単語番号（全語彙 V の各単語に 1 から $|V|$ までの番号を割り振ったもの）の要素のみが 1 で、それ以外の要素はすべて 0 とするベクトルである。次いで、隠れ層の出力と重み行列 $W'_{N \times |V|}$ を使って出力層への入力値を求める。

最後に、出力層には活性化関数を適応させる。活性化関数にはしきい値関数、シグモイド関数、Rectifier 関数等の種類があるが、word2vec では softmax 関数が利用される。

Softmax 関数を適用することで、予測対象単語 w の One-hot ベクトルが得られる。こうして得られた単語が目的の単語でなかった場合、つまり、ニューラルネットワークが間違った答えを出した場合、誤差を逆伝搬させて重み行列 $W_{|V| \times N}$、$W'_{N \times |V|}$ を更新する。

これまで述べてきたことを概念図として図 7.7 とて図 7.8 で示した。入力層と隠れ層の

間の重み行列 $W_{|V| \times N}$ の行ベクトル v_c が文脈単語 c の単語ベクトルで、隠れ層と出力層の間の重み行列 $W'_{N \times |V|}$ の列ベクトル $\tilde{v}_w$ が予測対象単語 w の単語ベクトルになる。

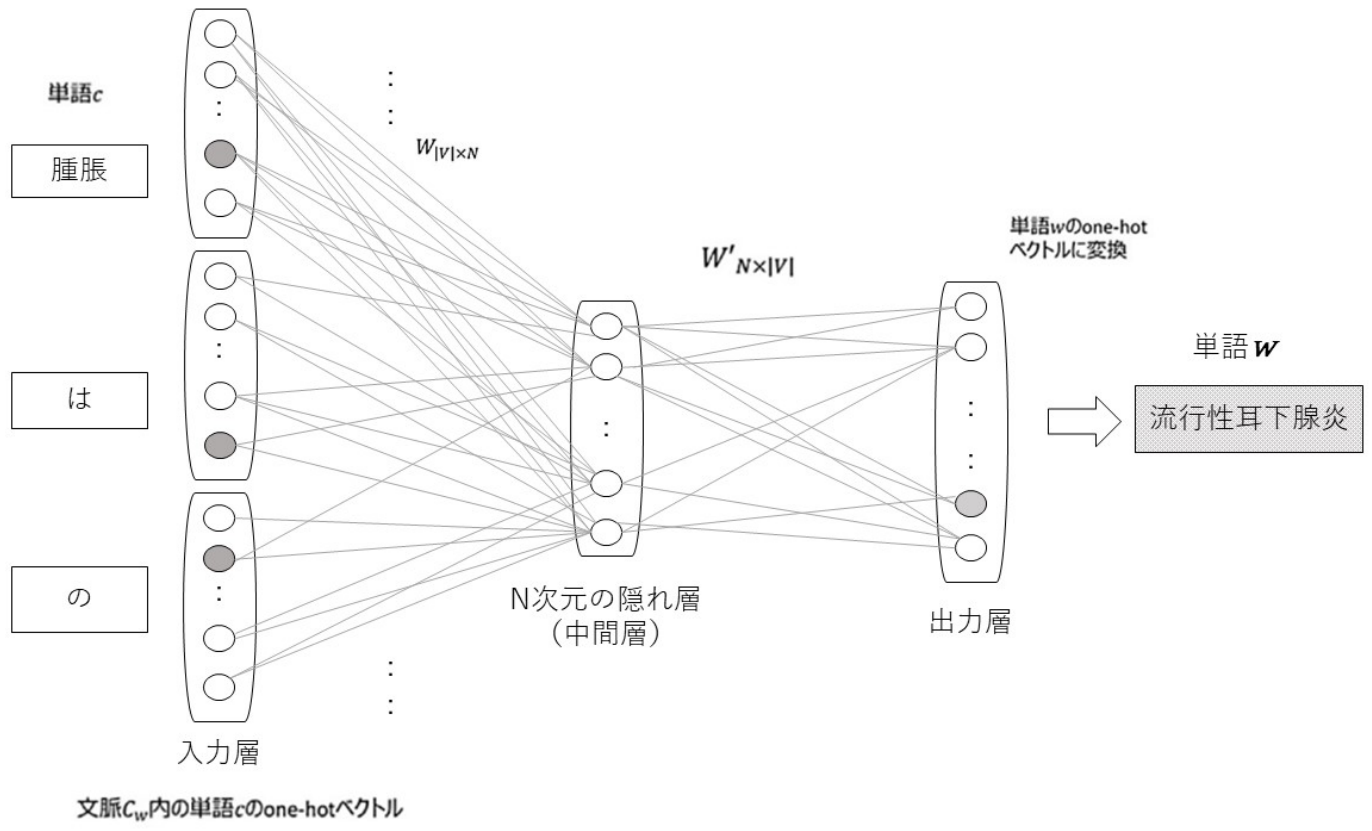


図 7.7　word2vec の skip-gram ニューラルネットワークモデル（田中 [32] の報告を元に筆者一部改編）

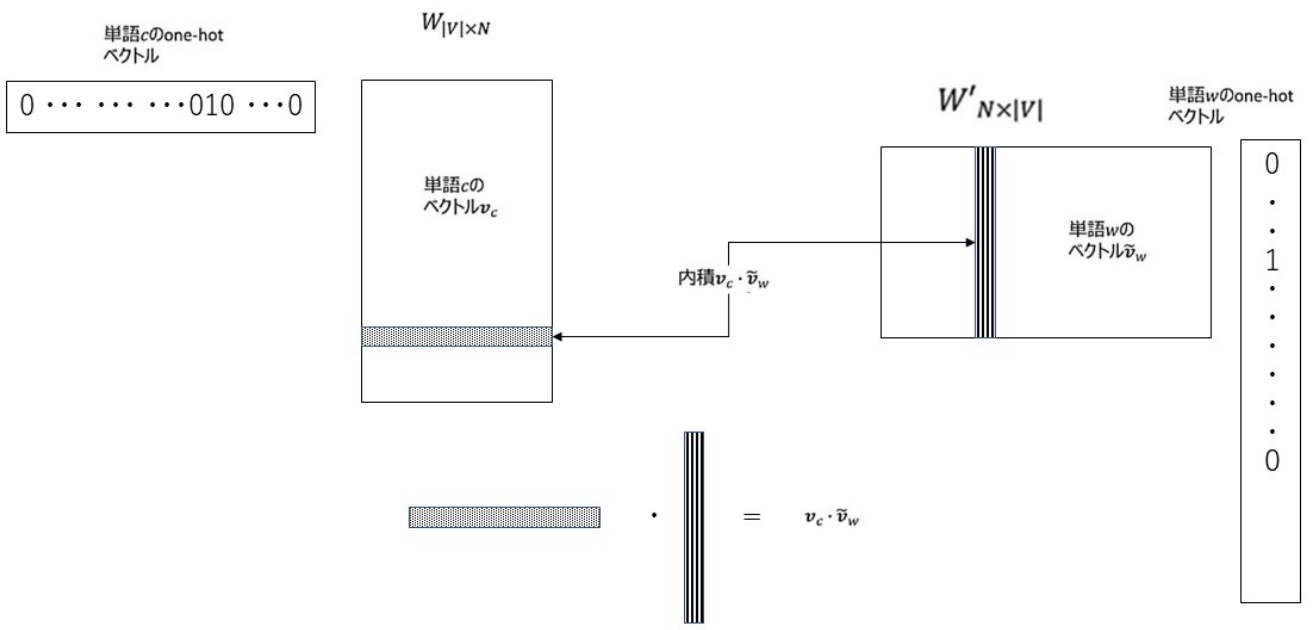


図 7.8　ニューラルネットワークと分散表現（田中 [32] の報告を元に筆者一部改編）

まとめると、Skip-gram では単語の重みベクトル同士の内積を計算していると見なせる。それが、出力層のユニットに入力される。ここで出力層への入力値に Softmax 関数を使うのは確立値に変換するからである。学習で使われているのは、ある単語とその単語に対して実際に現れる周辺語の内積が大きくなるように重みを調整していくといえる。

7.7　BERT

7.7.1　BERT

BERT とは Bidirectional Encoder Representations from Transformers の略称であり、2018 年に Google によって発表された自然言語処理モデルであり、再帰型ニューラルネットワークがベースになっている（図 7.9）[19, 33]。一つのトークンの出力を処理するに際、入力トークン全てを参照し、それらに重みをつけて出力の値を決定する。これによって文章を入力に与えた場合は前後の単語だけではなく文脈を考慮した出力が可能である。また、特徴として事前学習モデルであることが挙げられ、トークンの一部を隠して予測を行う Masked Language Model と Next Sentence Prediction という二つの文章がつながっているか否かを予測する方法の二つで学習が行われることが挙げられる。文脈を理解した分散表現が可能となる。

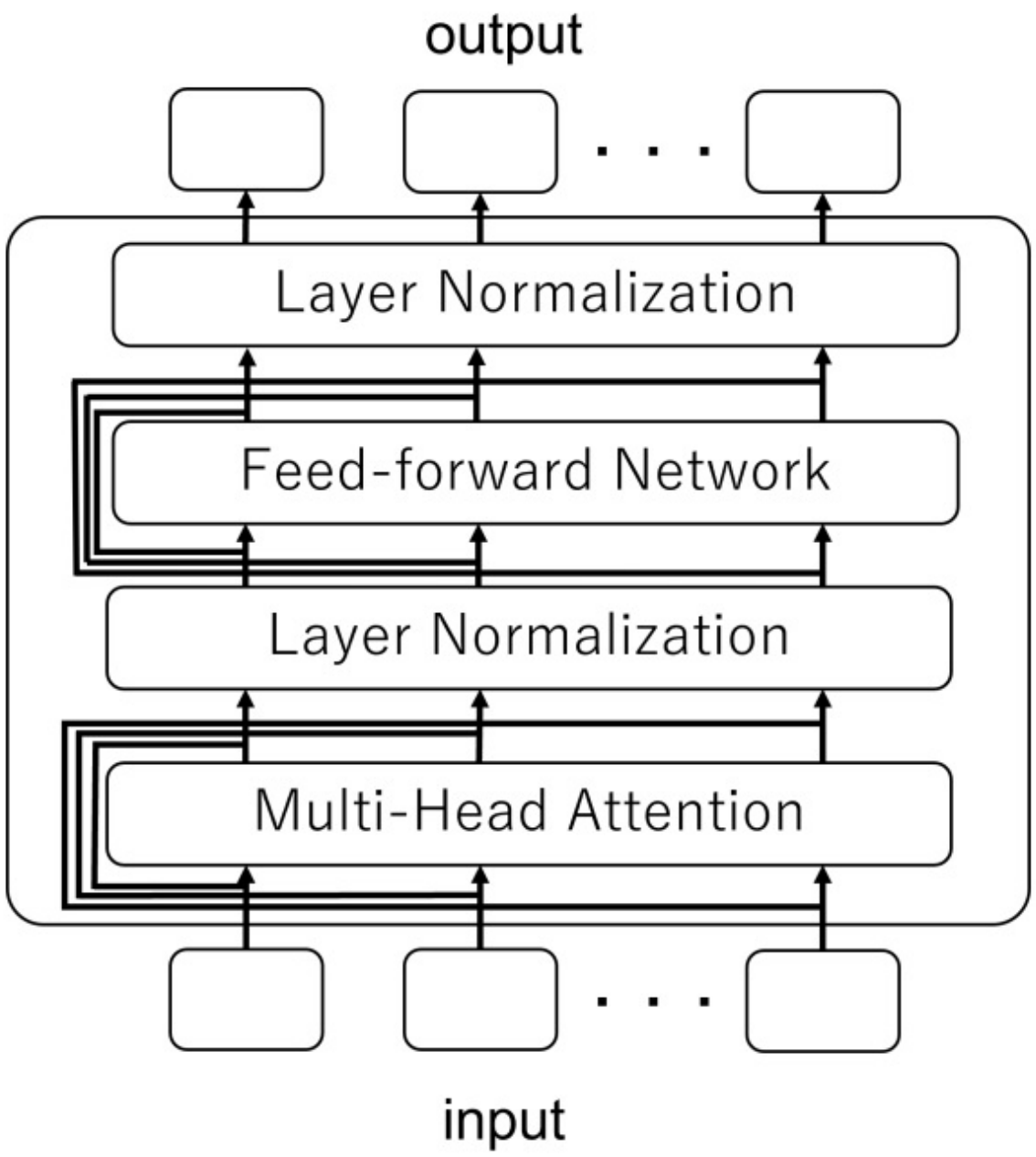


図 7.9　BERT の構造図　後藤貴樹のレポート [33] より抜粋

7.7.2　トークン化とベクトル化

BERT は複数の言語タスクに対応出来るような入力形式で設計されている。例として「明日は内視鏡の検査をする。」という文章は「明日」「は」「内視鏡」「の」「検査」「を」「する」「。」と言ったように分割される。そしてこのとき先頭と末尾に特殊なトークンが追加される。先頭に追加されるものを 「CLS」 トークンといい、末尾に追加されるものを 「SEP」 トークンという。「SEP」トークンは特殊トークンと呼ばれ、文章のペアの境界を示す役割や、入力の終わりを示す役割がある。「CLS」トークンも特殊トークンであり、「CLS」に対応する BERT の出力は、文章の分散表現として用いることができる。トークン列に分割した後は、それぞれのトークンをベクトルに置き換えて、BERT に入力する。トークン、文章のタイプ、文章中の位置それぞれに応じた三つのベクトルの和を BERT に入力する。

7.7.3　CLS

CLS トークンとは BERT によってトークン化された文章の先頭に配置されている特殊なトークンである。これに対する BERT の出力はトークンの分散表現ではなく文章そのものの分散表現としての使用が可能である。

7.7.4　SEP

SEP トークンは BERT によってトークン化された文章の末尾に配置される特殊なトークンである。これは末尾のみではなく、文章が二つ連続する場合も付加される。例として「今日は雨だ。家にいよう。」と言う文章なら「頭」「が」「痛」「い」「。」「SEP」「薬」「を」「のも」「う」「。」というように文章の継ぎ目にも配置される。

7.7.5　事前学習

事前学習は、大規模な文章コーパスを用いて汎用的な言語のパターンを学習するために行われる。学習と言うと「モデルに対して入力されるデータとそれに対する望ましい出力の関係を人間が付与したラベル付きデータを用いて学習させる」というイメージが一般的である。一方で、BERT の事前学習で用いるのは生の文章データのみである。このようにラベルが付与されていないデータをラベルなしデータと呼ぶ。

ラベルなしデータを用いるメリットは、比較的容易に大量のデータを収集できることである。例えば文章データはインターネットから容易に大量に収集できる。一方でラベルなしデータを用いた場合は、「モデルになにを学習させればよいか」ということがラベル付きデータを用いた学習と比べて明白ではなく、なんらかの処理が必要である。

BERT では「Masked Language Model」と「Next Sentence Prediction」の二つの手法を組み合わせて大量のデータから学習を行う。

・Masked Language Model：MLM

従来の次の単語を予想するタスクによって言語モデルを得る学習方法を用いると、双方型では予測する単語の情報を得ているため、学習がうまくいかない。そのため深い双方向性表現を訓練するために BERT では、入力トークンの何%かをランダムにマスクし、それらのマスクされたトークンを予測するという手法がとられている。この場合、標準的な手法と同様に、マスクトークンに対応する最終的な隠されたベクトルは、語彙を対象とした出力ソフトマックスに供給されることになる。この方法で双方向の事前学習モデルを得ることができる。

・Next Sentence Prediction：NSP

NLP タスクの多くが、文同士の関係性を理解することに基づいているということから、このタスクが設定された。関係性を学習するために、どのような単言語コーパスからでも簡単に生成できる二値化された次文予測タスクのための事前学習を行うという手法が行われている。

具体的には、各事前学習例の文 A と B を選択する際に、50 %の確率で B は A に続く実際の次の文（IsNext と表示）であり、50 %の確率でコーパスからのランダムな文（NotNext と表示）になるようにする。そして、次文なのか（IsNext か）、そうではないのか（NotNext か）を予測し、学習していく。

7.7.6　ファインチューニング

ファインチューニングとは特定のタスクに対する精度を向上させるための処理である。少数のラベル付き学習データを用いて BERT を学習させると共に分類器についても学習させることでその特定のタスクのみ精度が向上する。つまり言語タスクにおいて、BERT は特徴抽出器のような働きをする。BERT の出力を分類機等に接続するだけで、精度の高いモデルが作れることも特徴の一つである。

ファインチューニングを行うときは、モデルのパラメータの初期値として BERT のパラメータは事前学習で得られたパラメータを用い、新たに加えられた分類器のパラメータにはランダムな値を与える。そしてラベル付きデータを用いて、BERT と分類器の両方のパラメータを学習する。事前学習で得られたパラメータを初期値として扱うことで、比較的少数の学習データからでも高い性能のモデルを得ることができる。

第 8 章

データ分析

本研究で利用するデータは、富山大学附属病院で稼働している電子カルテの退院サマリーである。対象期間は 2004 年から 2019 年の 16 年分とした。富山大学附属病院は 2015 年に電子カルテのリプレースを行い、電子カルテシステム自体が変更（独自開発型電子カルテ NeoChart から大学病院向け電子カルテ EGMAIN-GX Enterprise Edition へリプレース）している。以下、リプレース前の電子カルテシステム（NeoChart）を旧電カル (図 8.1)、リプレース後の電子カルテシステム（EGMAIN-GX Enterprise Edition）を新電カル (図 8.2) とする。

この章ではこれらの退院サマリーのデータ分析を行う。機械学習の対象となるデータを抽出する際は、事前にデータクレンジング処理を施し対象データを抽出するため、実際のデータより少なくなるが、この章で示すデータは、登録されているデータをありのまま集計している点に留意されたい。

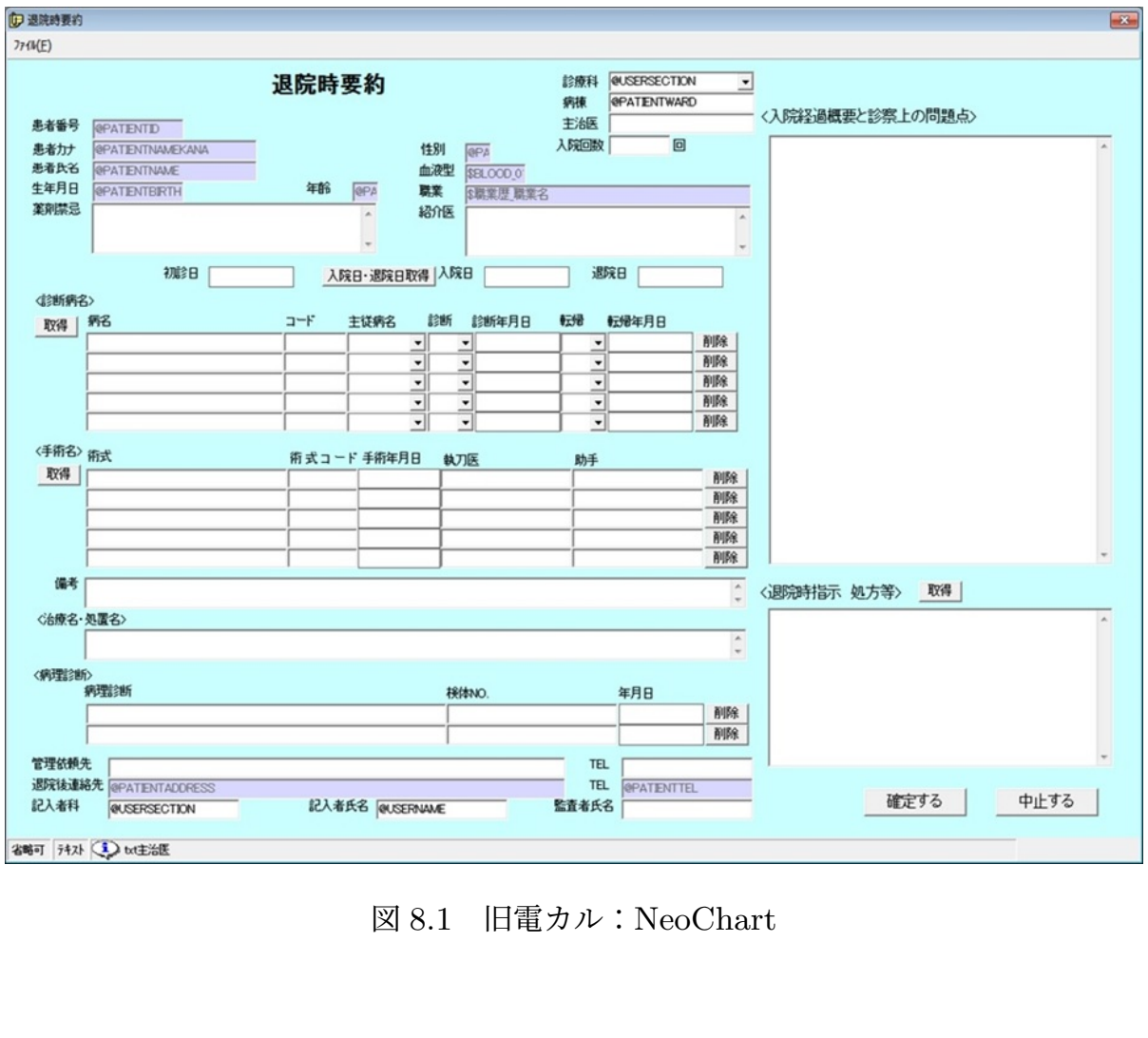


図 8.1　旧電カル：NeoChart

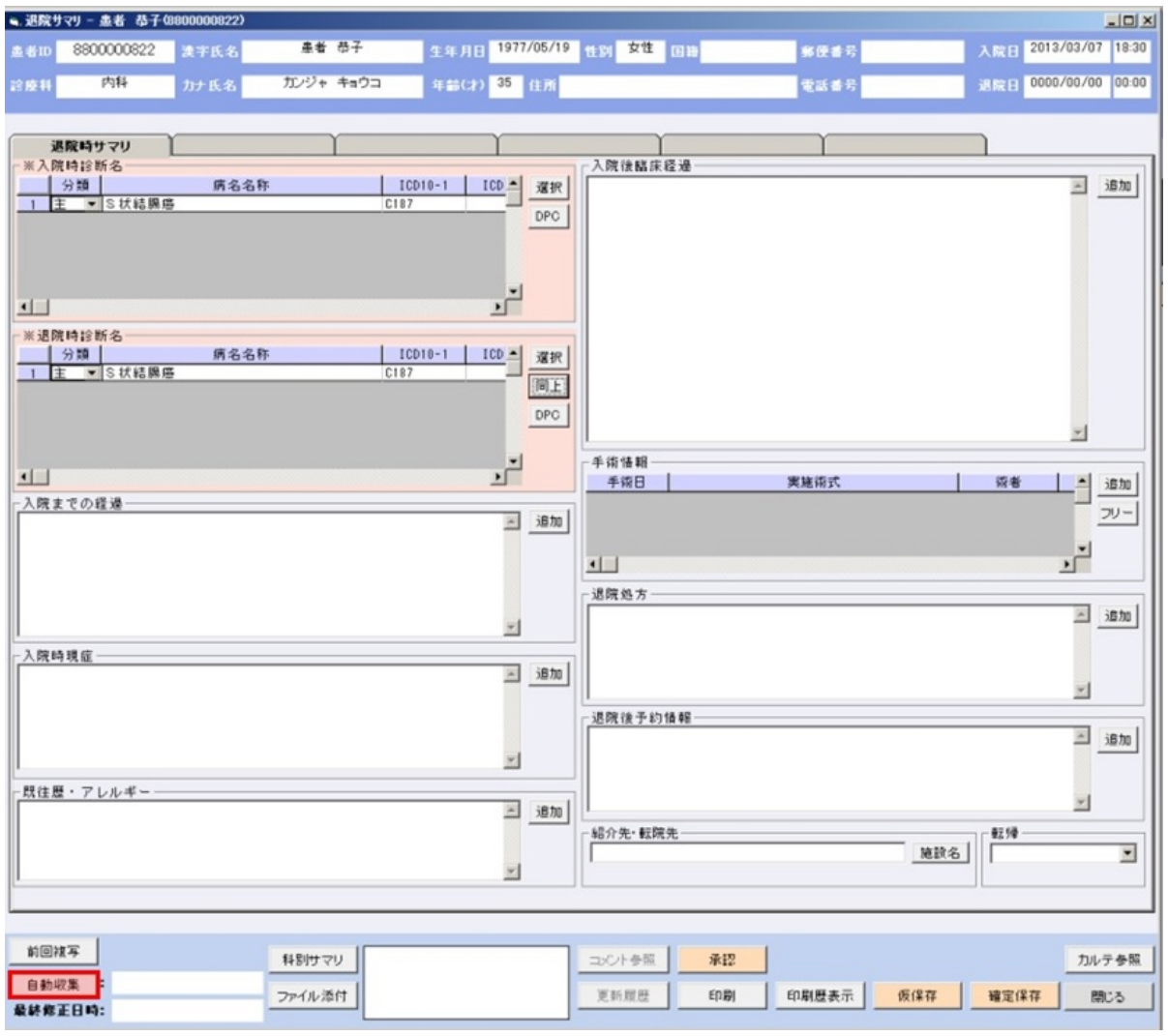


図 8.2　新電カル：EGMAIN-GX Enterprise Edition

8.1　対象データ数

対象データは旧電カルで入力された 94,083 件の退院サマリー、新電カルでは 61,772 件の退院サマリーを利用した。総計 155,855 件である。

8.2　病名数

旧電カルの総病名数は 3,204 病名、新電カルの総病名数は 2,849 病名であった。

8.3　年齢階層

旧電カル、新電カルの年齢階層ヒストグラムを図 8.3 に示す。

8.4　診療科毎登録件数

旧電カル、新電カルそれぞれの診療科毎登録件数を表 8.1 に示す。

表 8.1　新旧電カルの診療科毎退院サマリー件数

診療科名	旧電カル	新電カル
第一内科	8242	4803
第二内科	3175	5999
第三内科	11846	6102
第一外科	5602	3203
第二外科	7319	4152
口腔外科	2546	1668
和漢診療科	1264	160
小児科	5670	3138
感染症科	1	382
放射線科	310	N/A
整形外科	7067	3870
泌尿器科	6058	2928
救急科	593	951
産婦人科	11193	5913
皮膚科	2831	1963
眼科	8235	8201
神経内科	376	1410
精神神経科	2496	1286
総合診療部	22	183
耳鼻咽喉科	4597	2470
脳神経外科	4552	2639
麻酔科	88	12
臨床腫瘍部	N/A	339

8.5　経過要約文字数

旧電カル、新電カルそれぞれの経過要約文字数のヒストグラムを図 8.4 に示す。旧電カルの経過要約の総文字数は 139,601,677 文字、新電カルの経過要約の総文字数は 26,023,658 文字であった。

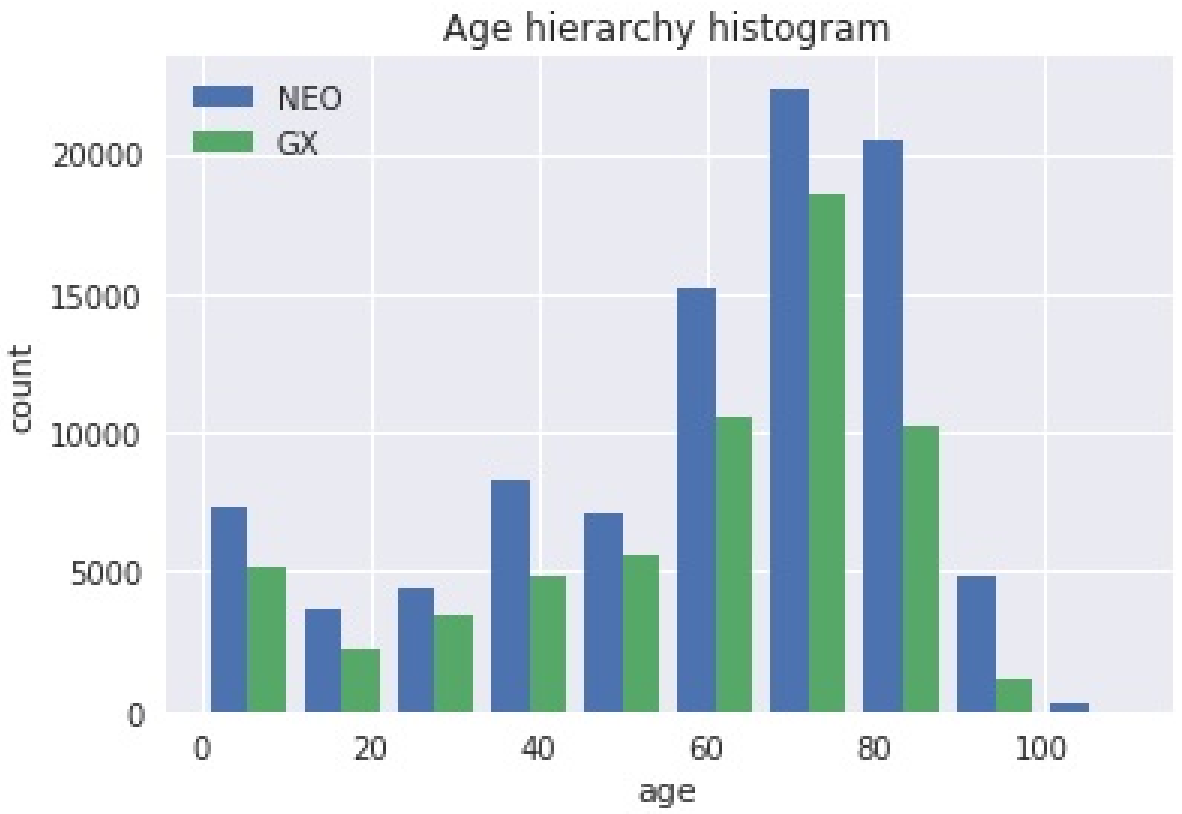


図 8.3　旧電カル (NEO)、新電カル (GX) の年齢階層ヒストグラム

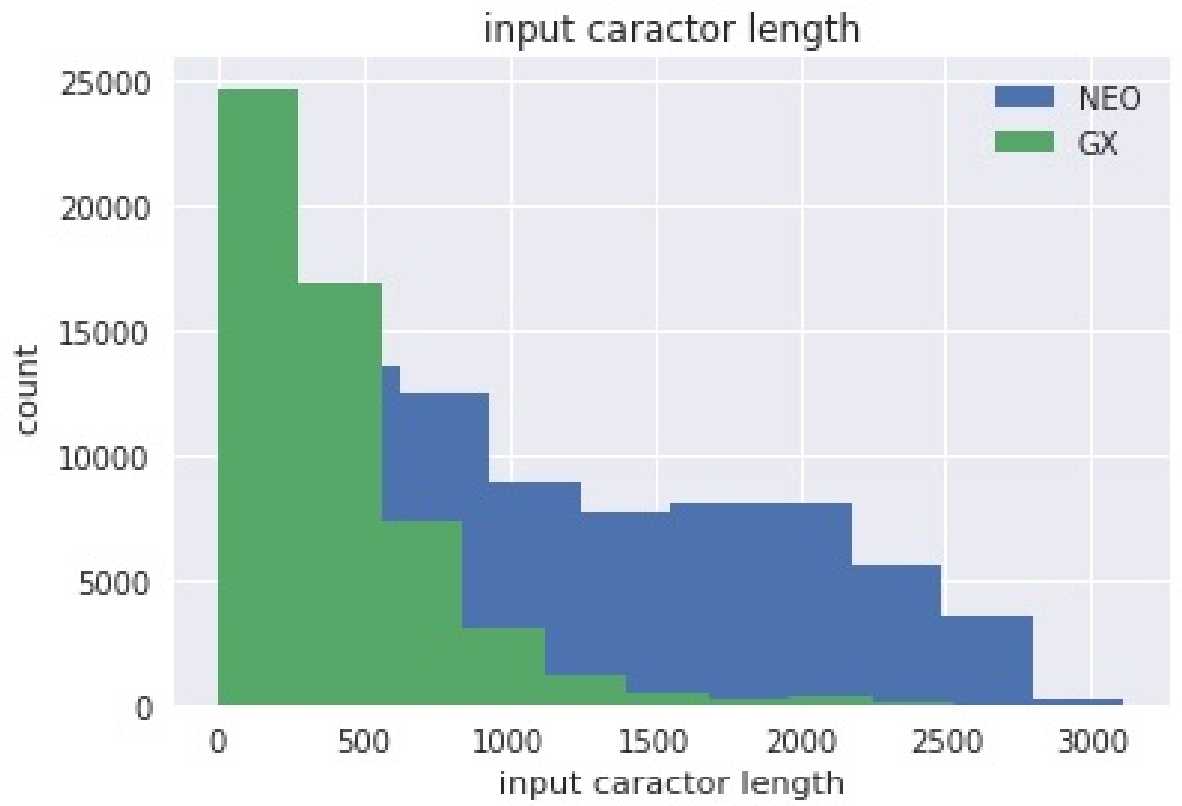


図 8.4　旧電カル (NEO)、新電カル (GX) の経過要約文字数の比較

8.6　登録件数上位 20 位診断病名コード

木村ら [30] の研究では、上位 20 件の病名を対象としている。CAC はメジャー病名を対象とすることから、本研究においても上位 20 病名を対象とした。

新電カル、旧電カルそれぞれの登録件数上位 20 位診断病名コードと登録件数を表 8.2、表 8.3 に示す。新旧電カルそれぞれにおいて上位 20 病名内に抽出された診断病名コードは、C22.0、C34.9、C56、C61、H33.0、H35.3 である（図 8.5)。なお、旧電カルでは手入力での病名入力が行われた時に便宜的に診断病名コードとして” 99999999”が割り当て

られていた。この手入力のコードの登録件数が 2,299 件あり、登録件数としては 1 位であったが、ICD10 コードが割り振られていないという理由から今回の順位表からは省いた。同様に新電カルでは診断病名コードとして“0”が割り当てられていたが、登録件数は 180 件であり上位 20 位以内には入っていない。

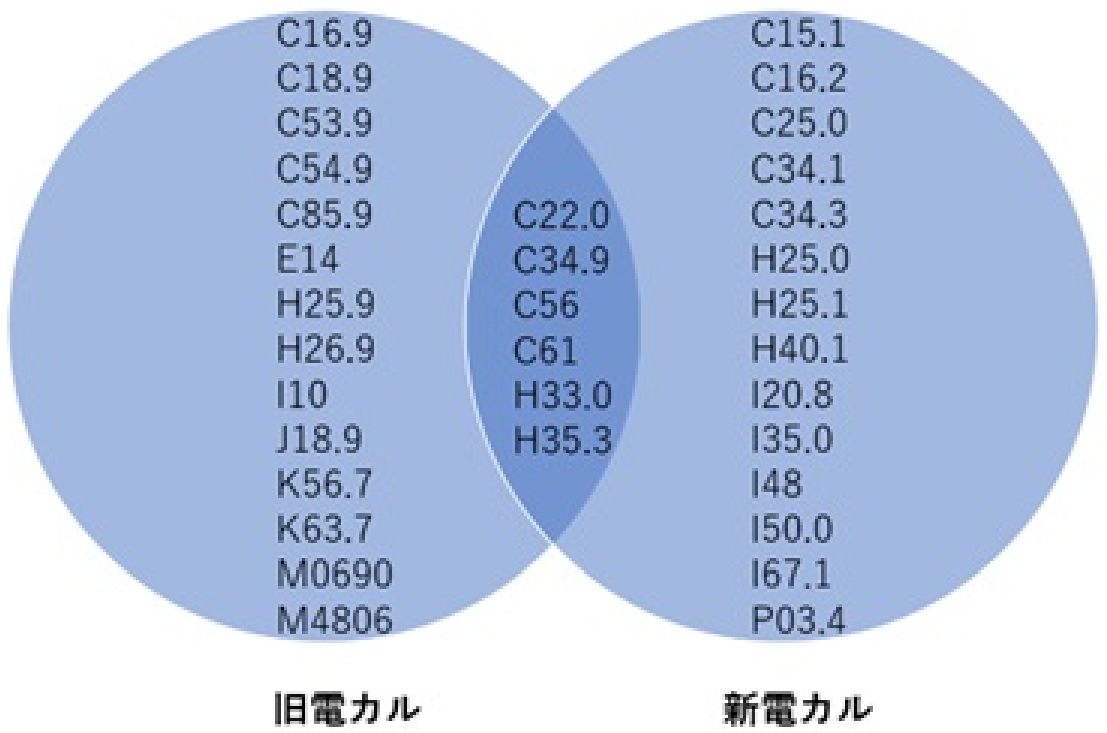


図 8.5　旧電カルと新電カル上位 20 位診断病名コードのベン図

新電カル上位 20 位診断病名コードに対応する旧電カルのレコード件数を表 8.4 に示す。旧電カルの上位 20 位の診断病名コードに対応した、旧電カルの診断病名コードが存在する事を確認した。

表 8.2　新電子カルの上位 20 位診断病名コード

診断病名コード	レコード件数
H25.1	2363
C34.1	1132
H35.3	975
C61	914
C34.3	897
C22.0	864
H40.1	798
H25.0	784
I20.8	700
I35.0	690
H33.0	616
I50.0	545
C16.2	536
I67.1	518
C25.0	503
I48	483
C15.1	483
C34.9	471
P03.4	437
C56	395

(新電カル:2015〜2019)

表 8.3　旧電子カルの上位 20 位診断病名コード

診断病名コード	レコード件数
C61	2278
H25.9	1640
C34.9	1589
C22.0	1503
C16.9	1333
C56	1276
H35.3	1104
M48.06	846
C53.9	809
K63.5	796
H26.9	757
C54.9	717
H33.0	629
I10	558
J18.9	552
M0690	552
K56.7	524
C85.9	511
C18.9	503
E14	499

(旧電カル:2004〜2014)

表 8.4　新電カル上位 20 位診断病名コードに対応する旧電カルのレコード件数

診断病名コード	新電カル	旧電カル
H25.1	2363	138
C34.1	1132	210
H35.3	975	1104
C61	914	2278
C34.3	897	158
C22.0	864	1503
H40.1	798	163
H25.0	794	323
I20.8	700	75
I35.0	690	70
H33.0	616	629
I50.0	545	166
C16.2	536	231
I67.1	518	387
C25.0	503	113
C15.1	483	212
I48	483	253
C34.9	471	1589
P03.4	437	400
C56	395	1276

第 9 章

提案手法

先行研究において、木村ら [9] は退院サマリーに対しシステム辞書により、形態素解析を行った。また、機械学習手法の決定木、RandomForest およびサポートベクターマシンを用いて実験を行い、サポートベクターマシンより RandomForest の正答率が高いと結論付けた。しかしながら、単純に形態素解析した単語の出現頻度によるデータセットを作成して機械学習を行っており解釈性に乏しい点、また形態素解析時にユーザー辞書として医療辞書を追加する事で正答率の向上が期待できる点で改善の余地はあるのではと考えた。

本研究では CAC の構築手法は機械学習を行う前に、意味表現学習と呼ばれるニューラルネットワークを用いた分散表現を利用し、経過要約の内容を「264 種類の特徴単語ベクトル値」に変換する。各特徴単語の重みを抽出することで、解釈性のある表現がベクトル値で表すことができることが本手法の優位性となる。機械学習の際には、経過要約に関しては自然言語で記載された記事は利用せず、「264 種類の特徴単語ベクトル値」である数値情報を利用する。さらに ICD10 病名と関連がある退院サマリの項目を説明変数として追加することで正答率が向上すると考えられ、堂坂ら [34] の研究を参考に、「264 種類の特徴単語」以外に、解釈性のある医療概念を含んだ説明変数として「診療科名」、「性別」、「年齢」を説明変数として追加した。

9.1　条件の変化による分析

手法の検討に際し、筆者はオオイらと共に旧電カルのデータセットを用い、条件の変化により正答率がどのように変化するかを調査した [40]。その結果を元に、本手法の提案根拠とする。

9.1.1　形態素解析で用いた辞書の変化による分析

形態素解析で用いた辞書はシステム辞書、NEologd 辞書および 2 種類の医療辞書（万病辞書、ComeJisyo）である。それぞれ形態素解析をする際に指定して、わかち書きを行った。図 9.1 と図 9.2 はシステム辞書と NEologd 辞書の比較である。レコードの中にある「脳挫傷」の経過要約の文字列「左上下肢筋緊張、右眼の軽度縮瞳、意識障害の遷

延」で形態素解析を行った。

```
keshi-lab-hp6@DESKTOP-KE8M3ND:~$ mecab
左上下肢筋緊張、右眼の軽度縮瞳、意識障害の遷延
左上    名詞,一般,*,*,*,*,左上,ヒダリウエ,ヒダリウエ
下肢    名詞,一般,*,*,*,*,下肢,カシ,カシ
筋      名詞,接尾,一般,*,*,*,筋,スジ,スジ
緊張    名詞,サ変接続,*,*,*,*,緊張,キンチョウ,キンチョー
、      記号,読点,*,*,*,*,、,、,、
右      名詞,一般,*,*,*,*,右,ミギ,ミギ
眼      名詞,一般,*,*,*,*,眼,メ,メ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
軽度    名詞,一般,*,*,*,*,軽度,ケイド,ケイド
縮      名詞,一般,*,*,*,*,縮,チヂミ,チジミ
瞳      名詞,一般,*,*,*,*,瞳,ヒトミ,ヒトミ
、      記号,読点,*,*,*,*,、,、,、
意識    名詞,サ変接続,*,*,*,*,意識,イシキ,イシキ
障害    名詞,一般,*,*,*,*,障害,ショウガイ,ショーガイ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
遷延    名詞,サ変接続,*,*,*,*,遷延,センエン,センエン
EOS
```

図 9.1　システム辞書

```
keshi-lab-hp6@DESKTOP-KE8M3ND:~$ mecab -d /usr/lib/mecab/dic/mecab-ipadic-neologd
左上下肢筋緊張、右眼の軽度縮瞳、意識障害の遷延
左上    名詞,一般,*,*,*,*,左上,ヒダリウエ,ヒダリウエ
下肢    名詞,一般,*,*,*,*,下肢,カシ,カシ
筋      名詞,接尾,一般,*,*,*,筋,スジ,スジ
緊張    名詞,サ変接続,*,*,*,*,緊張,キンチョウ,キンチョー
、      記号,読点,*,*,*,*,、,、,、
右      名詞,一般,*,*,*,*,右,ミギ,ミギ
眼      名詞,一般,*,*,*,*,眼,メ,メ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
軽度    名詞,一般,*,*,*,*,軽度,ケイド,ケイド
縮瞳    名詞,固有名詞,一般,*,*,*,縮瞳,シュクドウ,シュクドー
、      記号,読点,*,*,*,*,、,、,、
意識障害        名詞,固有名詞,一般,*,*,*,意識障害,イシキショウガイ,イシキショーガイ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
遷延    名詞,サ変接続,*,*,*,*,遷延,センエン,センエン
EOS
```

図 9.2　NEologd 辞書

また、医療辞書に入れ替え、上と同じ文字列で形態素解析を行った。

```
keshi-lab-hp6@DESKTOP-KE8M3ND:~$ mecab -u MANBYO_201907_Dic-utf8.dic
左上下肢筋緊張、右眼の軽度縮瞳、意識障害の遷延
左上    名詞,一般,*,*,*,*,左上,ヒダリウエ,ヒダリウエ
下肢筋緊張      名詞,サ変名詞,*,*,*,*,nan;icd=M6289;lv=E/freq=中頻度;筋強直,,,7
、      記号,読点,*,*,*,*,、,、,、
右      名詞,一般,*,*,*,*,右,ミギ,ミギ
眼      名詞,一般,*,*,*,*,眼,メ,メ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
軽度    名詞,形容動詞語幹,*,*,*,*,軽度,ケイド,ケイド
縮瞳    名詞,サ変名詞,*,*,*,*,しゅくどう;icd=H570;lv=S/freq=高頻度;縮瞳,しゅくどう,しゅくどう,210
、      記号,読点,*,*,*,*,、,、,、
意識障害        名詞,サ変名詞,*,*,*,*,いしきしょうがい;icd=R402;lv=S/freq=高頻度;意識障害,いしきしょうがい,いしきしょうがい,127383
の      助詞,連体化,*,*,*,*,の,ノ,ノ
遷延    名詞,サ変接続,*,*,*,*,遷延,センエン,センエン
EOS
```

図 9.3　万病辞書

```
keshi-lab-hp6@DESKTOP-KE8M3ND:~$ mecab -u ComeJisyoUtf8-1.dic
左上下肢筋緊張、右眼の軽度縮瞳、意識障害の遷延
左上下肢        名詞,一般,*,*,*,*,左上下肢,ヒダリジョウカシ,ヒダリジョウカシ,◆,:,,;,;,,:,看,:,3,:,41188
筋緊張  名詞,サ変接続,*,*,*,*,筋緊張,キンキンチョウ,キンキンチョウ,◆,:,muscle tone,;,,:,看助教,:,5,:,12168
、      記号,読点,*,*,*,*,、,、,、
右眼    名詞,一般,*,*,*,*,右眼,ミギメ,ミギメ,◆,:,,;,;,,:,看,:,4,:,40579
の      助詞,連体化,*,*,*,*,の,ノ,ノ
軽度    名詞,一般,*,*,*,*,軽度,ケイド,ケイド
縮瞳    名詞,一般,*,*,*,*,縮瞳,シュクドウ,シュクドウ,◆,:,miosis,;,,:,看教,:,2,:,16735
、      記号,読点,*,*,*,*,、,、,、
意識障害        名詞,一般,*,*,*,*,意識障害,イシキショウガイ,イシキショウガイ,◆,:,consciousness disturbance,;,,:,看栄教,:,4,:,7605
の      助詞,連体化,*,*,*,*,の,ノ,ノ
遷延    名詞,サ変接続,*,*,*,*,遷延,センエン,センエン
EOS
```

図 9.4　ComeJisyo

図 9.3 の万病辞書は「下肢筋緊張」と「縮瞳」をわかち書きしたことから、標準病名を含む形態素解析に適していることが分かる。また、図 9.4 の ComeJisyo は「左上下肢」と「右眼」をわかち書きし、医療現場で使われている言葉の形態素解析に適していることが確認できる。

9.1.2　説明変数の変化による分析

説明変数を増やすことにより正答率が向上すると考えられるため、「264 の特徴単語」以外、解釈性がある説明変数を追加した。まずは「診療科」の説明変数を追加する。図 9.5 の病名数を 20 件に絞った退院サマリの分析により、旧電子カルテ時代は 21 診療科が存在している。それぞれのレコードの診療科を One-hot 化し、重みの数値に近い値（0 か

1）とし、説明変数として追加した。

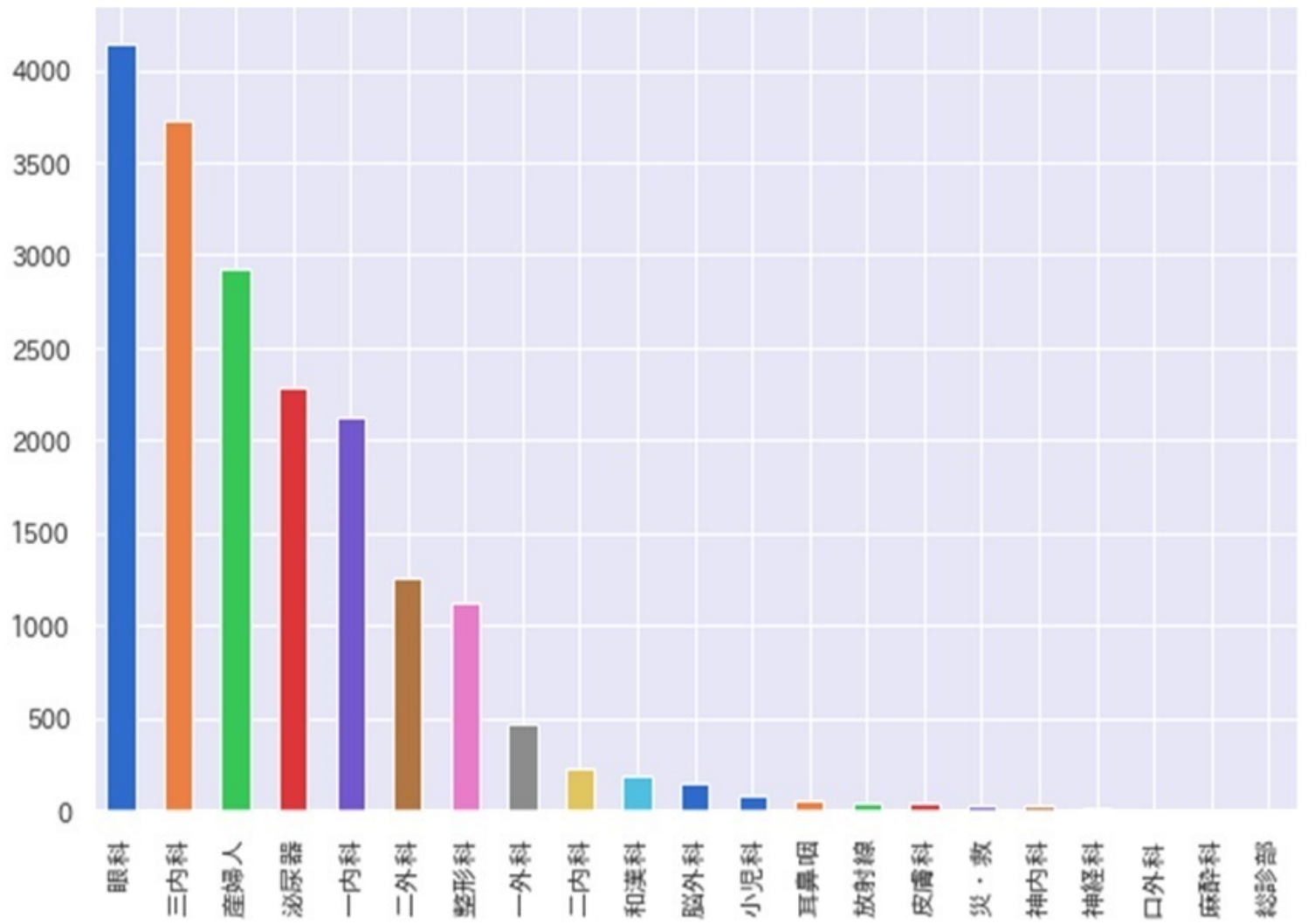


図 9.5　旧電カルにおける上位 20 病名における診療科毎の登録件数

加えて、「性別」の説明変数を追加する。図 9.6 は分析件数の 18,962 件に対して性別分けのグラフである。ここでも、それぞれの性別を One-hot 化して、重みの数値に近い値（0 か 1）の説明変数として追加し、機械学習を行った。

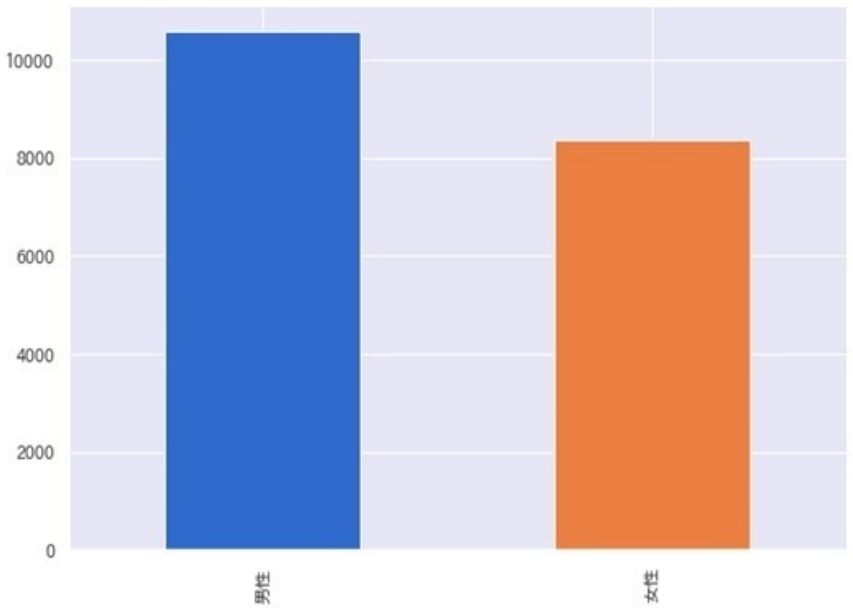


図 9.6　旧電カルにおける性別の数

さらに、「年齢」の説明変数を追加する。表 9.1 に 18,962 件に対する年齢の統計を示す通り、年齢は 1 から 106 まであり数値的には大きいため、それぞれの年齢を 100 で割って、特徴単語の重みに近い数値（0.01 から 1.06 まで）とし、説明変数に追加して、機械

学習を行った。

表 9.1　旧電カルにおける年齢の基本統計量

Count	18,962
Mean	72.152
Std	12.919
Min	1
25 %	65
50 %	74
75 %	81
Max	106

9.2　条件の変化による分析結果

9.2.1　形態素解析で用いた辞書の変化による分析結果

システム辞書に対して、NEologd 辞書は生活上よく使う言葉が多く含まれているため、NEologd システム辞書を用いて解析を行った。表 9.2 は学習から分析までの設定条件である。また、表 9.3 はそれに対して機械学習を行った時の 64 種類の分析病名を目的変数とした時の正答率の結果である。

表 9.2　NEologd 辞書の設定条件

使用辞書	Wiki コーパス	学習回数	学習文字数	学習件数	分析病名数	分析件数	目的変数	説明変数
NEologd	なし	1	500	72,623	68	33,177	病名コード	264 特徴単語

表 9.3　NEologd 辞書の分析結果

決定木（%）	RandomForest（%）	線形 SVM（%）	非線形 SVM（%）
24.8	35.0	68.8	75.8

また、医療辞書の万病辞書と ComeJisyo を指定して解析を行った。表 9.4 は学習から分析までの設定条件である。表 9.5 はそれに対して機械学習を行った結果である。サポートベクタマシンの結果は NEologd 辞書より医療辞書の方が正答率が高く、医療辞書の利用は病名の解釈性があると考えられた。

表 9.4　医療辞書の設定条件

使用辞書	Wiki コーパス	学習回数	学習文字数	学習件数	分析病名数	分析件数	目的変数	説明変数
万病辞書＆ ComeJisyo	なし	1	500	72,623	68	33,177	病名コード	264 特徴単語

表 9.5　医療辞書の分析結果

決定木（%）	RandomForest（%）	線形 SVM（%）	非線形 SVM（%）
25.3	35.1	69.2	76.1

9.2.2　説明変数の変化による分析結果

木村ら [9] は病名数を 20 件に絞って、機械学習を行うことで高い正答率を求めた。そこで本研究においても主要な病名を対象としているため、同条件である上位 20 病名を対象（表 9.6 の設定条件）としたところ、表 9.7 の結果が得られた。病名数を 20 件に絞ることで、サポートベクタマシンの正答率が 10 %上がった。

表 9.6　病名数絞りの設定条件

使用辞書	Wiki コーパス	学習回数	学習文字数	学習件数	**分析病名数**	分析件数	目的変数	説明変数
万病辞書＆ ComeJisyo	なし	1	500	72,623	**20**	18,962	病名コード	264 特徴単語

表 9.7　病名数絞りの分析結果

決定木（%）	RandomForest（%）	線形 SVM（%）	非線形 SVM（%）
49.9	56.7	82.5	86.1

さらに、医療辞書の利用と病名数 20 個の基準条件とし、表 9.8 に示すように、One-hot 化した診療科名を説明変数に追加した。説明変数が 264 個から 285 個とし、機械学習を行った。その結果を表 9.9 に示す。

表 9.8　説明変数の追加による設定条件

使用辞書	Wiki コーパス	学習回数	学習文字数	学習件数	分析病名数	分析件数	目的変数	**説明変数**
万病辞書＆ ComeJisyo	なし	1	500	72,623	20	18,962	病名コード	**264** 特徴単語 **21** 診療科

表 9.9　説明変数の追加による分析結果

決定木（%）	RandomForest（%）	線形 SVM（%）	非線形 SVM（%）
60.9	63.1	81.2	86.7

また、表 9.10 に示すように、One-hot 化した性別を説明変数に追加した。ここでは、説明変数が 285 個から 287 個となり、機械学習を行った。その結果を表 9.11 に示す。

表 9.10　性別を追加した設定条件

使用辞書	Wiki コーパス	学習回数	学習文字数	学習件数	分析病名数	分析件数	目的変数	**説明変数**
万病辞書＆ ComeJisyo	なし	1	500	72,623	20	18,962	病名コード	**264** 特徴単語 **21** 診療科 **2** 種類の性別

表 9.11　性別を追加した分析結果

決定木（%）	RandomForest（%）	線形 SVM（%）	非線形 SVM（%）
61.0	62.0	83.8	86.8

これらの分析結果より、条件として、退院サマリに登録された上位 20 病名を対象とし、経過要約の先頭 500 文字を意味表現学習を用いて 264 種類の特徴単語ベクトル値に変換する。さらに「診療科」＋「年齢」＋「性別」を説明変数として追加したものを、非線形サポートベクタマシン（One versus the rest）にて機械学習を行った時が最も正答率が高かったため、本研究では、CAC 構築手法として「意味表現学習＋非線形サポートベクタマシン」を提案することとした。

第 10 章

CAC 実装

10.1　データセット作成フロー

機械学習を行うまでに作成する、意味表現学習を利用したデータセットの作成フローを図 10.1 に示す。

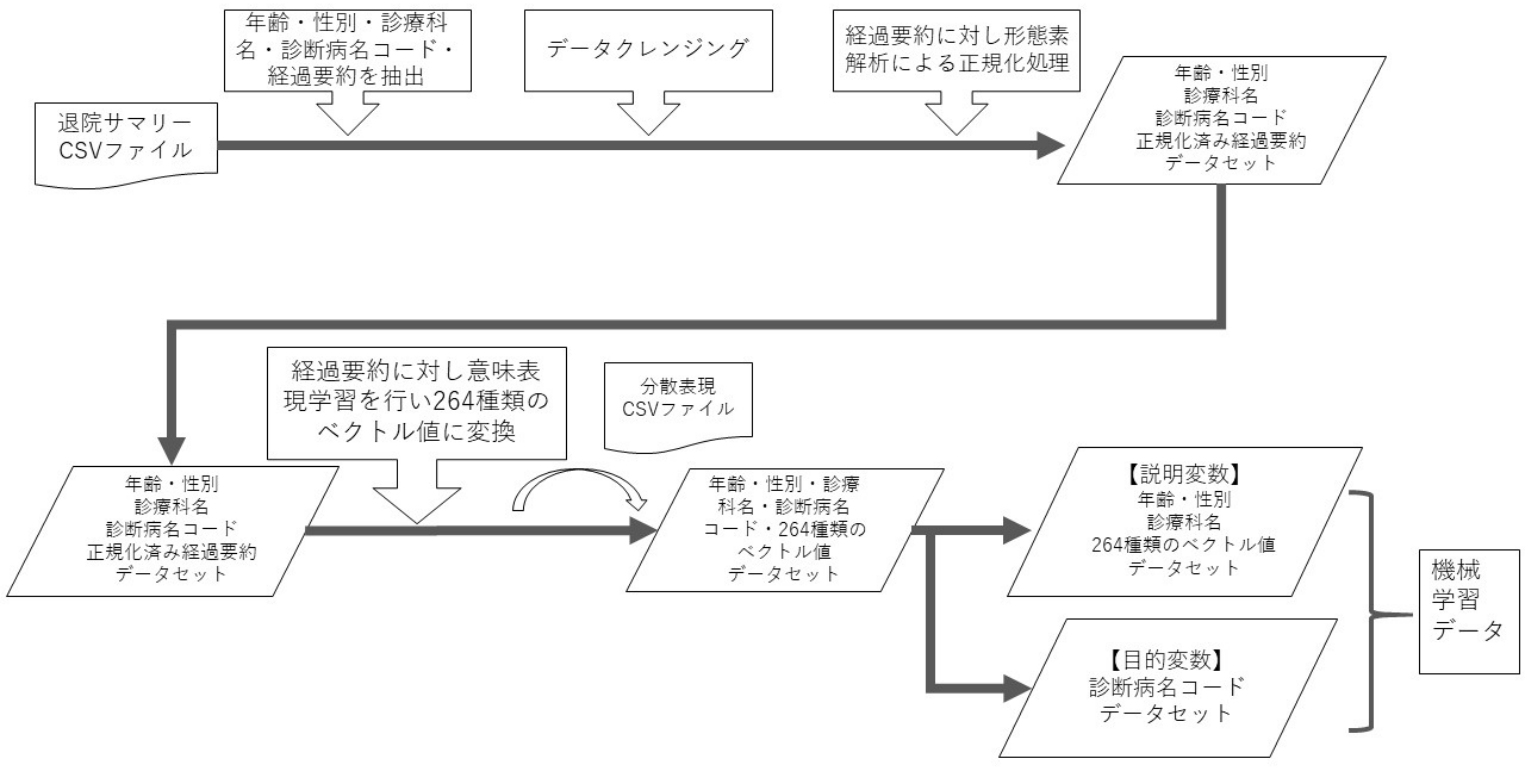


図 10.1　電子カルテの退院サマリーから機械学習用データセットを作るまでの流れ

10.2　データクレンジング

CAC の構築には機械学習の処理を行う。機械学習のための教師データの作成のためにデータクレンジング処理をした。教師データ作成のために欠落値を省く処理を行った。欠落値を省く対象は利用する変数、すなわち説明変数で用いる、「経過要約」、「性別」、「年齢」、「診療科名」と目的変数で用いる「診断病名コード」とした。さらに、k 匿名化処理を施すため、総レコード数の 0.02 %以下の登録件数しかない病名を削除した。また、経過要約の入力文字数が 50 未満のレコードに対して追加で除外処理を行った。データクレンジング後の総退院サマリー数は旧電カルが 73,150 件、新電カルが 49,611 件であった。データクレンジング後の旧電カルの上位 20 位までの診断病名コードを表 10.1、新電カルの上位 20 位までの診断病名コードを表 10.2 に示す。新旧それぞれにランクインした診断病名コードは、C61、C34.9、C22.0、C56、H35.3、M48.06、H33.0 である（図 10.2)。データクレンジング後の新電カルの上位 20 位までの診断病名コードに対応するデータクレンジング後の旧電カルのレコード件数を表 10.3 に示す。データクレンジング後も新電カルの上位 20 位までの診断病名コードに対応した経過要約が旧電カルに存在する事を確認した。

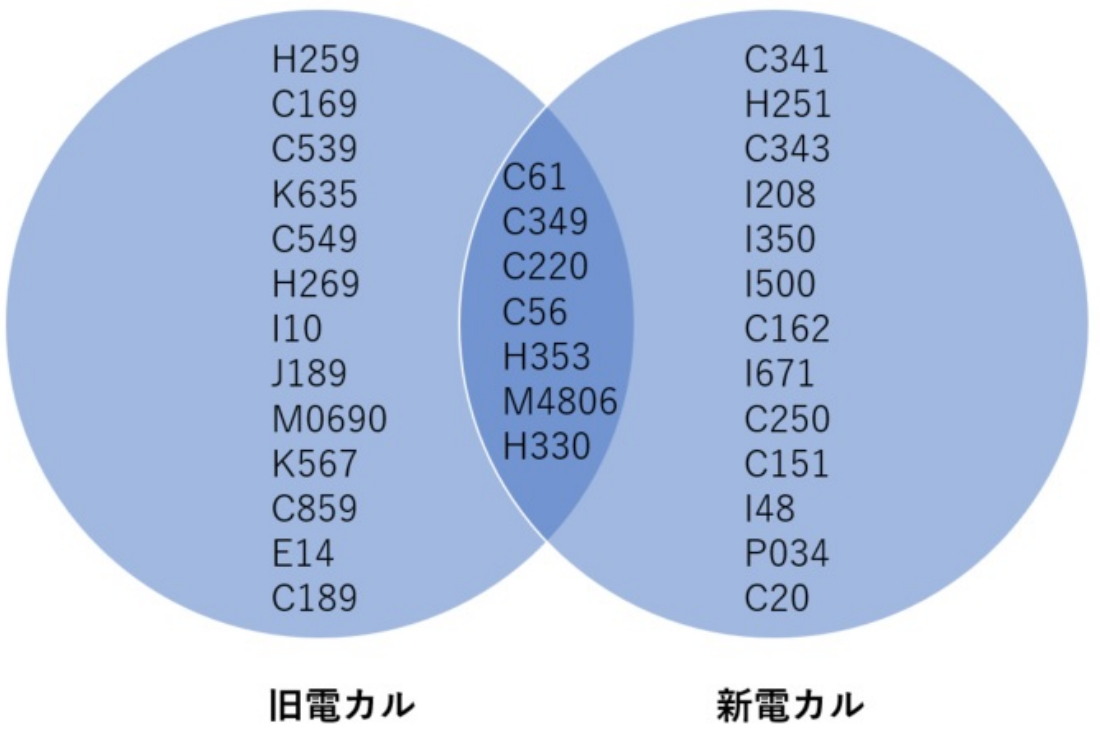


図 10.2　データクレンジング後の旧電カルと新電カル上位 20 位診断病名コードのベン図

表 10.1　新電カルデータクレンジング後の上位 20 位診断病名コード

診断病名コード	レコード件数
C34.1	1127
H25.1	929
C61	912
C34.3	893
C22.0	864
I20.8	698
I35.0	690
I50.0	545
C16.2	536
I67.1	515
C25.0	503
C15.1	483
I48	483
C34.9	468
P03.4	432
C56	393
M48.06	373
H35.3	368
H33.0	361
C20	357

(新電カル:2015〜2019)

表 10.2　旧電カルデータクレンジング後の上位 20 位診断病名コード

診断病名コード	レコード件数
C61	2216
H34.9	1579
C34.9	1566
C22.0	1501
C16.9	1323
C56	1276
H35.3	1060
M48.06	845
C53.9	807
K63.5	775
C54.9	716
H26.9	671
H33.0	625
I10	558
J18.9	552
M0690	539
K56.7	521
C85.9	506
E14	499
C18.9	498

(旧電カル:2004〜2014)

表 10.3　新電カルデータクレンジング後の上位 20 位診断病名コードに対応する旧電カルデータクレンジング後のレコード件数

診断病名コード	新電カル	旧電カル
C34.1	1127	210
H25.1	929	123
C61	912	2216
C34.3	893	158
C22.0	864	1501
I20.8	698	75
I35.0	690	70
I50.0	545	166
C16.2	536	231
I67.1	515	387
C25.0	503	111
C15.1	483	211
I48	483	253
C34.9	468	1579
P03.4	432	399
C56	393	1276
M48.06	373	845
H35.3	368	1060
H33.0	361	625
C20	357	343

10.3　形態素解析

「経過要約」に形態素解析による分かち書きを行った。形態素解析にはMeCabを利用し、医療辞書として万病辞書とComeJisyoを用いた。分かち書きの際に病名の用語を標準病名に変換して出力し、表記揺れの問題を改善した。その他の品詞に対しては原型を出力した。また単語のアルファベットと数字、カタカナに関しては半角、特殊記号は除く正規化処理を行った。形態素解析後の経過要約の総文字数は旧電カルが29,883,189文字、新電カルが14,874,929文字であった。

10.4　意味表現学習

形態素解析後の「経過要約」に意味表現学習を行い、264種類の特徴単語に対する重みを抽出した。意味表現学習とは自然言語で記載された非構造化のテキストを、約2万語の単語に関係ある特徴単語を列挙した辞書（単語意味ベクトル辞書）を初期値としたニューラルネットワークを用いて、人が解釈可能な分散表現を求める学習方法である[36, 37, 38]。特徴単語の264種類は百科事典の概念分類の264種類に対応する。そのため、人が解釈可能な分散表現を求めることが可能となる。

図10.3は老人性白内障の経過要約の文字列を意味表現学習して、重みの高い特徴単語を10個表現した例である。経過要約の特徴単語と特徴単語の関連性は「視力低下」＝「暗さ」、「受診」＝「医学・薬学」、「左目」＝「人間の体」と考えられる。

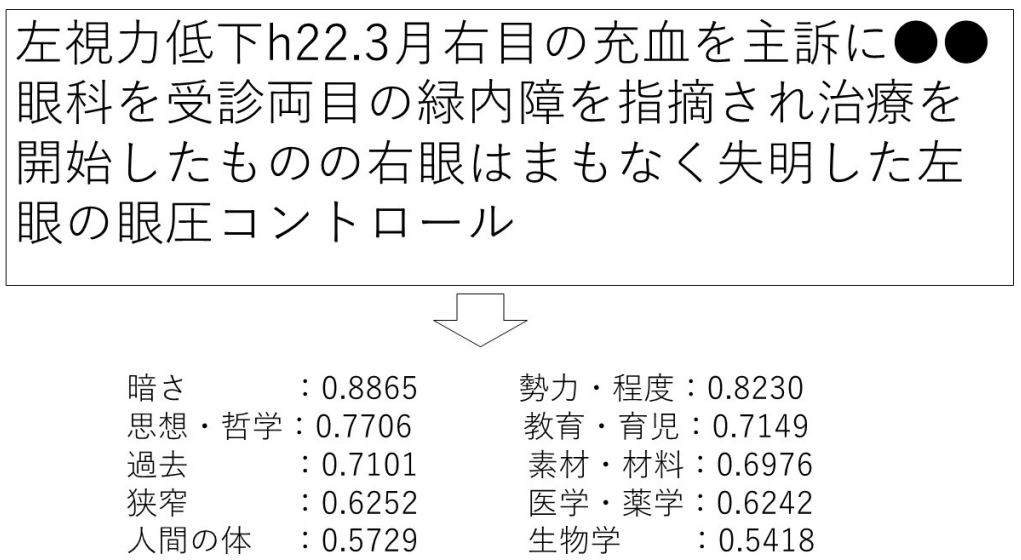
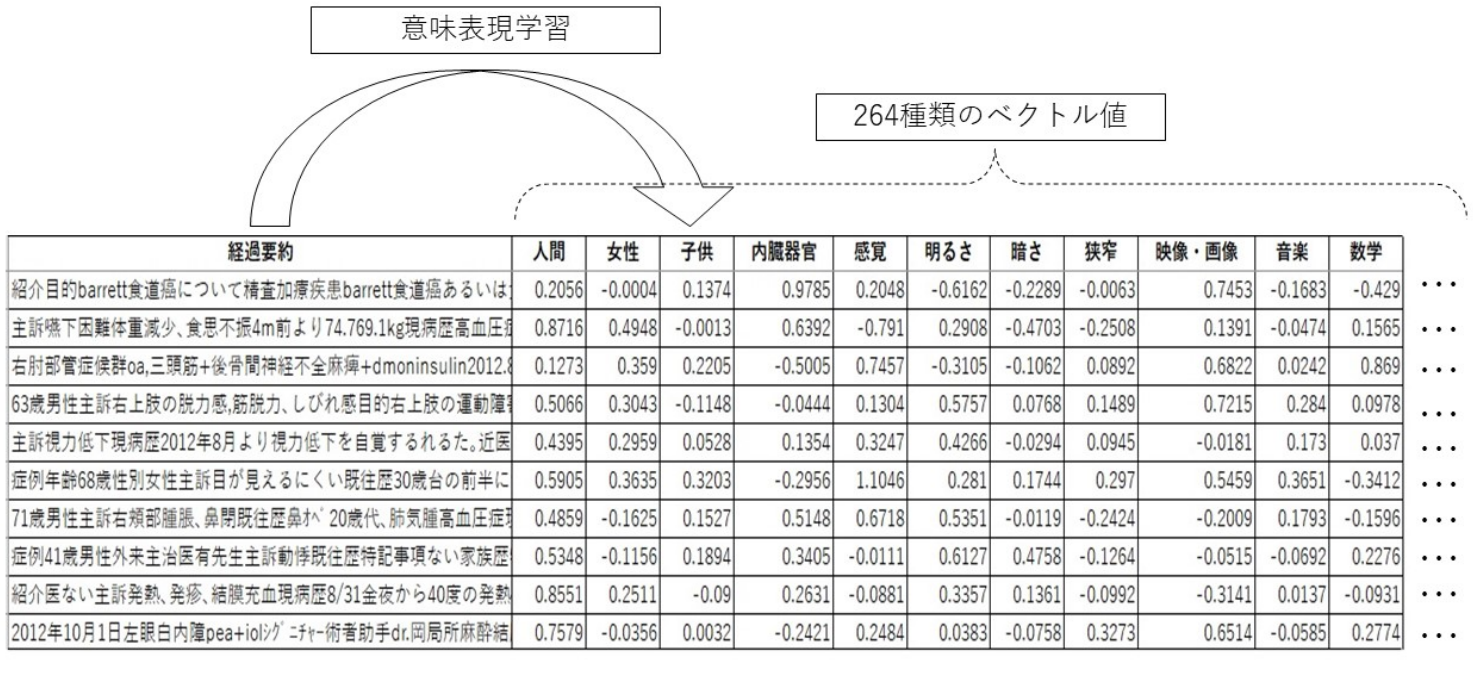


図10.3　経過記録を意味表現学習した例

経過要約の先頭 500 文字を対象に意味表現学習を行った。特徴単語数は百科事典の概念分類 264 種類とし、各経過要約に対応した特徴単語ベクトル値を求めた（図 10.4）。意味表現学習の処理手順を以下に示す。



経過要約	人間	女性	子供	内臓器官	感覚	明るさ	暗さ	狭窄	映像・画像	音楽	数学	
紹介目的barrett食道癌について精査加療疾患barrett食道癌あるいは	0.2056	-0.0004	0.1374	0.9785	0.2048	-0.6162	-0.2289	-0.0063	0.7453	-0.1683	-0.429	・・・
主訴嚥下困難体重減少、食思不振4m前より74.769.1kg現病歴高血圧症	0.8716	0.4948	-0.0013	0.6392	-0.791	0.2908	-0.4703	-0.2508	0.1391	-0.0474	0.1565	・・・
右肘部管症候群oa,三頭筋+後骨間神経不全麻痺+dmoninsulin2012.8	0.1273	0.359	0.2205	-0.5005	0.7457	-0.3105	-0.1062	0.0892	0.6822	0.0242	0.869	・・・
63歳男性主訴右上肢の脱力感,筋脱力、しびれ感目的右上肢の運動障害	0.5066	0.3043	-0.1148	-0.0444	0.1304	0.5757	0.0768	0.1489	0.7215	0.284	0.0978	・・・
主訴視力低下現病歴2012年8月より視力低下を自覚するれるた。近医	0.4395	0.2959	0.0528	0.1354	0.3247	0.4266	-0.0294	0.0945	-0.0181	0.173	0.037	・・・
症例年齢68歳性別女性主訴目が見えるにくい既往歴30歳台の前半に	0.5905	0.3635	0.3203	-0.2956	1.1046	0.281	0.1744	0.297	0.5459	0.3651	-0.3412	・・・
71歳男性主訴右頬部腫脹、鼻閉既往歴鼻ﾎﾟ20歳代、肺気腫高血圧症現	0.4859	-0.1625	0.1527	0.5148	0.6718	0.5351	-0.0119	-0.2424	-0.2009	0.1793	-0.1596	・・・
症例41歳男性外来主治医有先生主訴動悸既往歴特記事項ない家族歴	0.5348	-0.1156	0.1894	0.3405	-0.0111	0.6127	0.4758	-0.1264	-0.0515	-0.0692	0.2276	・・・
紹介医ない主訴発熱、発疹、結膜充血現病歴8/31金夜から40度の発熱	0.8551	0.2511	-0.09	0.2631	-0.0881	0.3357	0.1361	-0.0992	-0.3141	0.0137	-0.0931	・・・
2012年10月1日左眼白内障pea+iolｼﾞﾆﾁｬｰ術者助手dr.岡局所麻酔結	0.7579	-0.0356	0.0032	-0.2421	0.2484	0.0383	-0.0758	0.3273	0.6514	-0.0585	0.2774	・・・

図 10.4　意味表現学習を行い、経過要約の内容を「264 種類の特徴単語ベクトル値」に変換した。

10.4.1　語彙辞書作成

経過要約では語彙辞書を作成する時に以下の 2 種類の初期ベクトルを最初に構築する。

・264 種類の特徴単語を語彙として追加し、その初期ベクトルを特徴単語に対応する次元を 1 とする 264 次元の One-hot ベクトルとする。

・264 種類の特徴単語を除く、経過要約から抽出された全単語（経過要約に含まれる基本単語を含む）の初期ベクトルは、264 次元のゼロベクトルとする。

語彙辞書の各単語は、入力ベクトルと出力ベクトルの 2 種類のベクトルを持つ。入力ベクトルは入力層の単語と隠れ層の 264 種類のノードとの間の重みであり、出力ベクトルは出力層の単語と隠れ層の 264 種類のノードとの間の重みに対応する。入力ベクトル、出力ベクトル共に上記の初期化を行った。

10.4.2　シードベクトル作成

辞書の単語に対する定義文を再帰展開することにより単語ベクトルを生成する手法は提案されており、単語意味ベクトル辞書は、264 種類の特徴単語で基本単語を定義しているとみなすことが出来る。特徴単語も基本単語となるため再帰展開が必要だが、定義文が 264 語に限定されるため、数回展開すれば収束する。本研究では，Faruqui[39] らのレトロフィッティングツールを用いて、単語意味ベクトル辞書を再帰的に展開すること

により、基本単語のシードベクトルを生成した（図 10.5）。

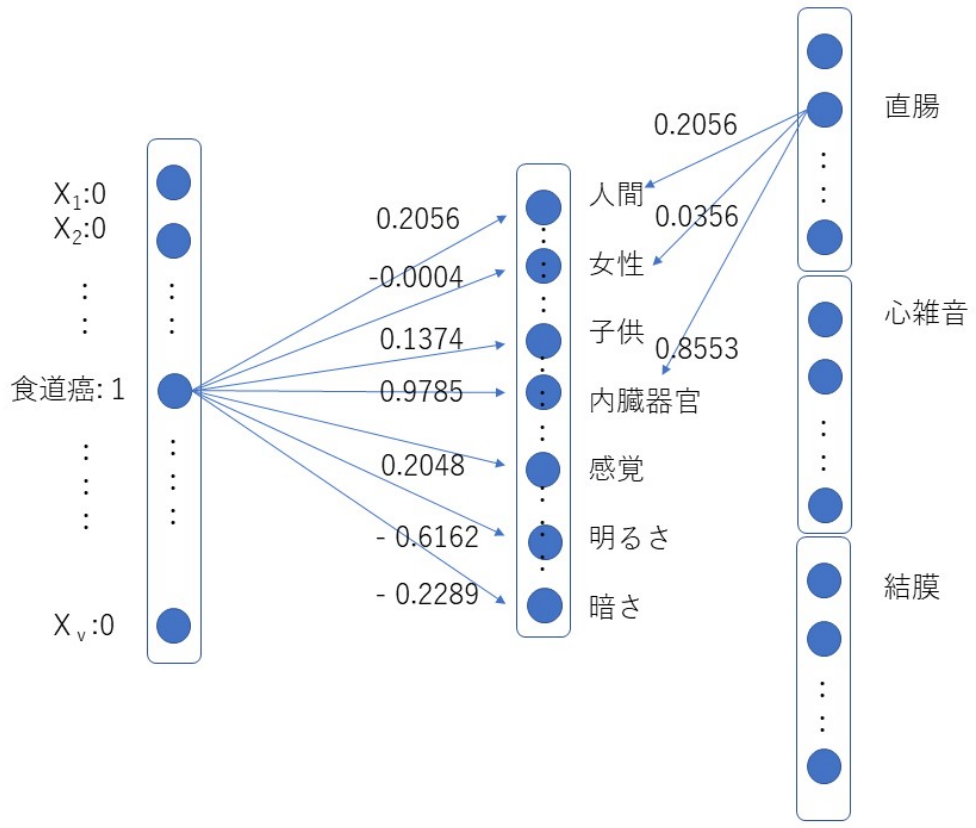


図 10.5　Skip-gram モデルへのシードベクトルの設定

10.4.3　単語ベクトル作成

単語意味ベクトル辞書は約 2 万語と少ないため、この基本単語の入力・出力ベクトルを Skip-gram モデルのシードベクトルとして、今回作成した旧電カルと新電カル、それぞれの経過要約コーパスを用いて、単語ベクトルの学習を行う。学習回数は 3 回とした。

10.4.4　パラグラフベクトル作成

分析対象となる経過要約コーパスを用いて、パラグラフベクトルの学習を行う。シードベクトルには 10.4.3 で学習した単語ベクトルを用いた。学習回数は 20 回とした。

10.4.5　分散表現の CSV ファイル出力

学習結果により得られた経過要約毎の 288 種類の特徴単語の重みを CSV ファイルとして出力した。

10.5　説明変数と目的変数の設定

説明変数を「264 種類の特徴単語のベクトル値」、「性別」、「年齢」、「診療科名」とし、目的変数を「診断病名コード」とした。「性別」と「診療科名」はカテゴリ変数、「264 種類の特徴単語のベクトル値」と「年齢」は数値型変数として扱った。なお、標準退院サマリーの規約では「診療科名」は必須項目では無いが、記載者情報として主治医情報を

記載することになっている。本提案手法は標準退院サマリーの導入を見据えて構築する意図があり、主治医情報と関連性が強い項目として「診療科名」を説明変数として追加した（図 10.6）。

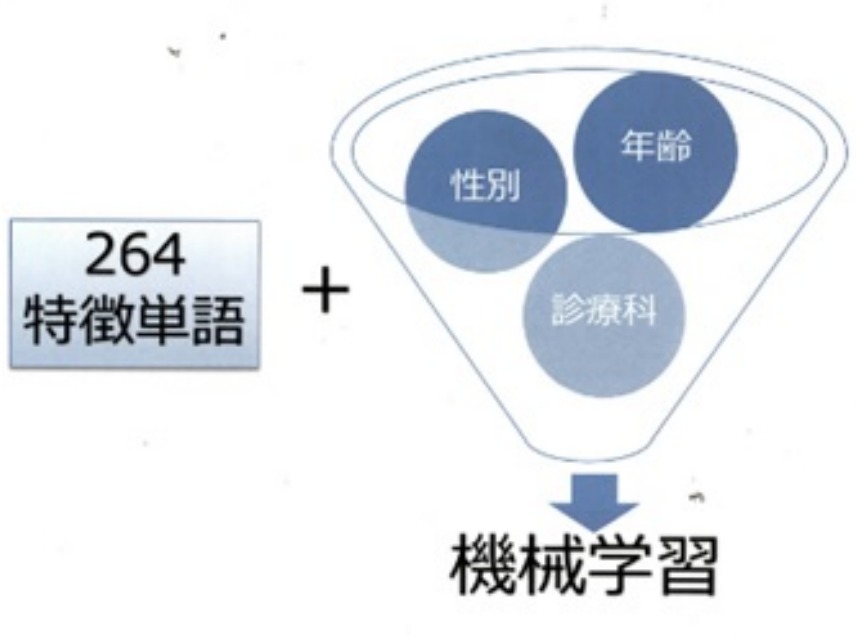


図 10.6　264 の特徴単語以外に、「性別」「年齢」「診療科名」を説明変数に追加した.

10.6　訓練用データセットと評価用データセットの作成

機械学習では、訓練用データセットから学習モデルを作る。その学習モデルを評価用データセットに当てはめて性能を評価する。評価する診断病名コードは、新電カルの上位 20 位までの診断病名コードとする。そのため訓練用データセット、評価用データセット共に新電カル上位 20 位の診断病名コードを含んだレコードに絞った。通常、訓練用データセットと評価用データセットの割合は 7:3 もしくは 8:2 を用いることが多いが、今回は、データセット新電カルの上位 20 位までの病名で絞り込んでいるため、旧電カルにおいては表 7 にあるように I20.8 や I35.0 のように登録件数が少ないものもある。そのため訓練用データセットと評価用データセットの割合がほぼ同等数になっていることに留意されたい。

10.7　機械学習による 3 種類の評価方法の目的と意味

検証を目的に３種類の学習モデルを用意した。それぞれ評価１、評価２、評価３とする。

評価１ではモデルの分布が異なるデータセットにおいても汎用的に性能を発揮できるかの確認を目的とした。旧電カルのクレンジング後の 73,150 件のコーパス、新電カルのクレンジング後の 49,611 件のコーパスを用い、新旧のコーパスでそれぞれ意味表現学習を行う。意味表現学習後のファイルより新電カル上位 20 病名を抽出し、訓練用と評価用

データセットを作り、旧電カルデータで学習モデルを作成し、新電カルデータで評価する。訓練用データセットは 11,839 件、評価用データセットは 11,930 件である（図 10.7）。

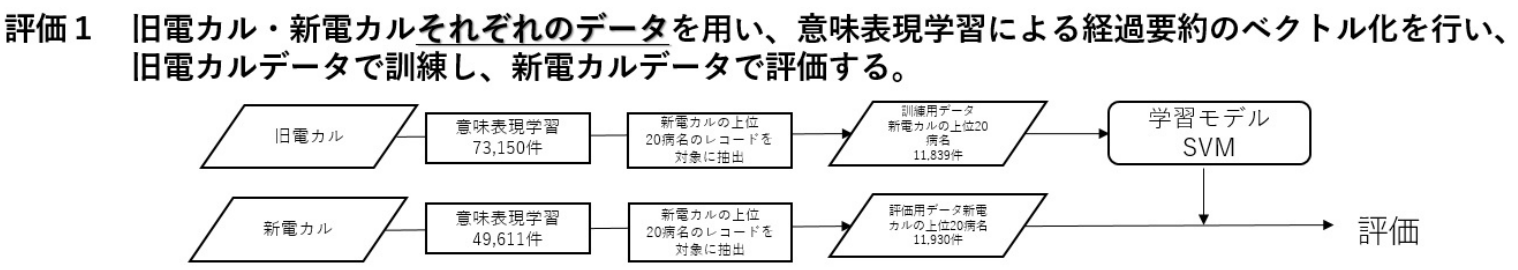


図 10.7　評価１の処理フロー

評価 2 では、意味表現学習の対象レコードを拡張し評価 1 との差異を測定する目的とした。旧電カルのクレンジング後の 73,150 件と新電カルのクレンジング後の 49,611 件の 2 つのコーパスを併せたファイルを用意し、総数 122,761 件のコーパスを用いて意味表現学習を行った。意味表現学習後のファイルより新電カル上位 20 病名を抽出し、訓練用と評価用データセットを作り、旧電カルデータで学習モデルを作成し、新電カルデータで評価する。訓練用データセットは評価 1 と同じく 11,839 件、評価用データセットは 11,930 件である（図 10.8）。

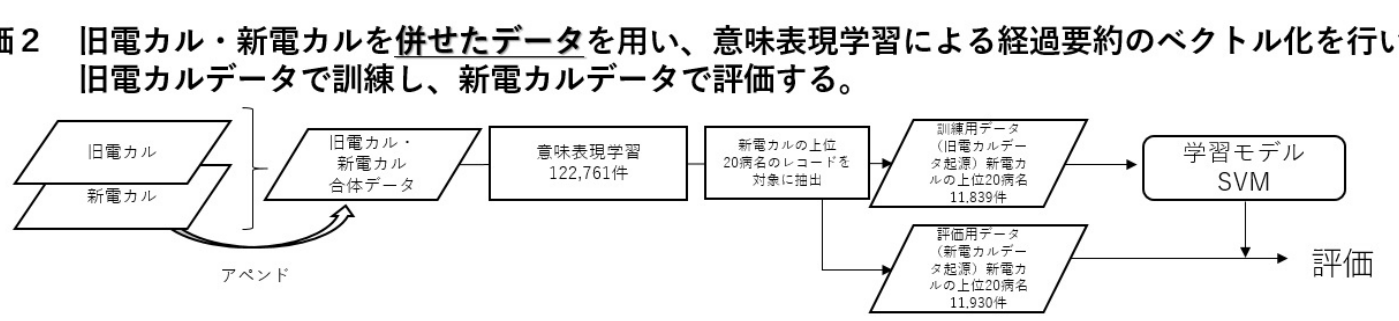


図 10.8　評価 2 の処理フロー

評価 3 では、評価 1 の結果で実施した「意味表現学習＋サポートベクタマシン」の組み合わせが CAC として安定的に動作するかを目的とした。新電カルのクレンジング後の 49,611 件のコーパスを用いて意味表現学習を行う。意味表現学習後のファイルより新電カル上位 20 病名を抽出し、評価用と訓練用データセットを作り、新電カルのデータを利用して訓練用データセットと評価用データセットを作成する。この場合のデータセットの分割はシステムの規定値である 75:25 を採用した（図 10.9）。

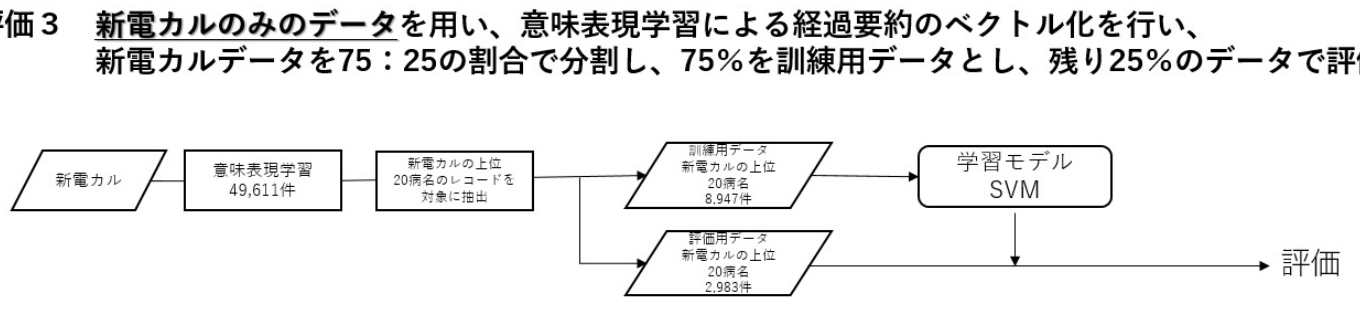


図 10.9　評価 3 の処理フロー

第 11 章

性能評価

まず評価 1 ではデータクレンジング後の新旧それぞれのデータに対し意味表現学習を行う。意味表現学習を行った後、新電カル上位 20 病名のレコードのみを抽出する。旧電カルの訓練用データセットを用いて学習モデルを作成し、新電カルで作成した評価用データセットに当てはめて学習モデルの検証を行う。評価対象の診断病名コードは表 10.1 に示したデータクレンジング後の新電カル上位 20 位の診断病名コードとした。

学習モデルの検証には、各診断病名コードの適合率（Precision）と再現率（Recall）を求める。この時、適合率と再現率はトレードオフの関係にあり双方を同時に伸ばすことはできない。この 2 つの指標をまとめるものとして、適合率と再現率との調和平均を意味する F 値（図 11.1）がある。適合率と再現率をバランスのとれた大きさを求めるには、F 値を最大にすることで達成できる。先行事例では正答率を求めているが、正答率の場合、誤りの要因が捉えにくいため、本研究では各クラス内のより詳細な要因の把握が可能となる F 値で機械学習におけるクラス分類の評価を行った。

次に 20 の診断病名コードが対象のため 20 クラスを表現する混合行列を作成し予測結果を表で確認した。混合行列の説明を図 11.2 に示す。

例えばC 15.1 が正解である 300 レコードがある場合、正しく予測されたのが 240 回、C16.2 と予測したのが 20 回、C20 と予測したのが 40 回であった場合の再現率は上記式より 80 %である。また適合率は 88.9 %となる。F 値は上記 F1 の式より 84.2 %と求まる。マクロ F 値とはそれぞれの F 値の平均である。この図右下の 78.4 %がマクロ F 値である。

機械学習は線形サポートベクターマシン、RBF カーネルを利用した非線形サポートベクターマシンを採用し、それぞれの F 値を比較した。非線形サポートベクターマシンは事前にグリッドサーチによるハイパーパラメータの検索を行い、ベストパラメータにて実行している。評価 1 の機械学習の結果の比較表を表 11.1 に示す。

線形サポートベクターマシンのマクロ F 値は 0.536 であり、非線形サポートベクターマシンの F 値は 0.595 であった。非線形サポートベクターマシンの方がやや良い評価を得れたため、今回は非線形サポートベクターマシンを採用した。非線形サポートベクターマシンの混合行列を求め表 11.2 に示した。

混合行列

予測クラス

実際クラス	Positive	Negative
Positive	True Positive	False Negative → 再現率 Recall
Negative	False Positive ↓ 適合率 Precision	True Negative

$$\text{recall} = \frac{TP}{TP+FN} \qquad \text{precision} = \frac{TP}{TP+FP} \qquad F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

図 11.1　適合率と再現率と F 値の関係

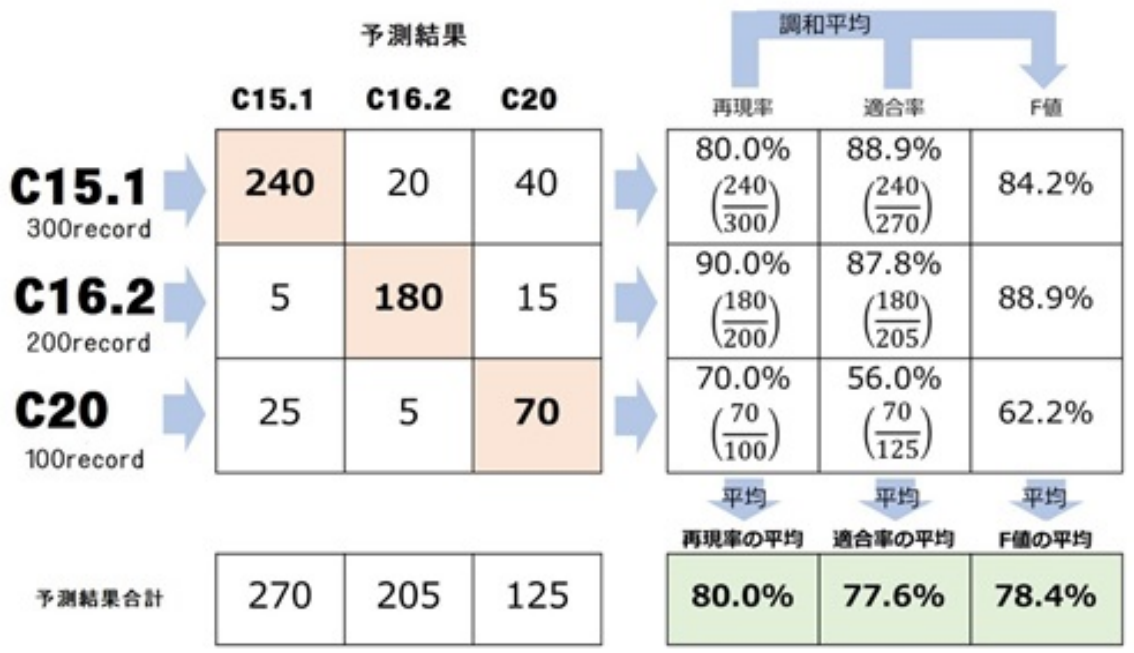


図 11.2　多クラス分類の場合の混合行列の説明

表 11.1　旧電カルをモデルとした新電カル上位 20 位診断病名コードの各機械学習の F 値（適合率と再現率の調和平均）の比較表

診断病名コード	線形 SVM	非線形 SVM	評価件数
		C=100, gamma=0.1	
C15.1	0.707	0.760	483
C16.2	0.625	0.745	536
C20	0.274	0.401	357
C22.0	0.755	0.797	864
C25.0	0.142	0.273	503
C34.1	0.247	0.165	1127
C34.3	0.113	0.007	893
C34.9	0.335	0.332	468
C56	0.863	0.966	393
C61	0.949	0.975	912
H25.1	0.116	0.551	929
H33.0	0.493	0.602	361
H35.3	0.333	0.255	368
I20.8	0.687	0.758	698
I35.0	0.141	0.213	690
I48	0.599	0.591	483
I50.0	0.547	0.584	545
I67.1	0.931	0.974	515
48.06	0.904	0.959	373
P03.4	0.961	0.994	432
マクロ平均 F 値	0.536	0.595	11930

表 11.2　評価 1 の混合行列

	C15.1	C16.2	C20	C22.0	C25.0	C34.1	C34.3	C34.9	C56	C61	H25.1	H33.0	H35.3	I20.8	I35.0	I48	I50.0	I67.1	M48.06	P03.4
C15.1	336	67	4	28	1	0	0	28	6	5	0	0	0	0	0	2	5	1	0	0
C16.2	47	322	7	86	0	1	0	18	10	22	0	1	0	5	0	9	3	1	1	3
C20	29	88	69	131	0	0	0	6	8	1	0	0	0	1	1	10	3	0	4	6
C22.0	0	0	0	854	1	2	0	0	0	0	0	0	0	0	0	1	6	0	0	0
C25.0	51	14	49	240	39	3	0	14	30	3	0	3	0	0	0	9	30	1	9	8
C34.1	2	1	4	6	2	189	47	822	27	2	0	1	0	0	0	0	6	17	1	0
C34.3	0	1	3	7	0	115	57	663	11	6	0	0	0	1	0	1	12	9	5	2
C34.9	0	0	5	0	3	25	6	413	7	0	1	0	3	0	0	1	0	2	2	0
C56	0	0	4	1	0	0	0	1	380	2	0	1	0	0	0	0	1	1	0	2
C61	0	0	0	0	0	0	0	1	0	908	0	0	0	0	0	0	3	0	0	0
H25.1	0	0	0	0	0	0	0	0	0	0	60	195	672	1	0	0	0	1	0	0
H33.0	0	0	0	0	0	0	0	0	0	0	8	219	133	0	0	0	0	0	0	1
H35.3	0	0	0	0	0	0	0	0	0	0	27	105	235	0	0	0	0	1	0	0
I20.8	1	0	0	30	0	13	5	7	0	12	4	0	0	440	1	70	67	13	35	0
I35.0	0	0	1	5	1	48	3	14	0	22	3	1	0	89	53	241	182	18	8	1
I48	1	0	0	8	0	2	0	0	0	10	0	0	0	6	5	404	39	5	1	2
I50.0	0	2	1	2	1	3	1	13	3	7	0	1	0	38	1	117	340	3	11	1
I67.1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	512	0	0
M48.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	371	0
P03.4	0	0	0	0	0	0	0	0	5	0	0	0	0	2	0	1	0	0	0	424

次に、評価 2 としてデータクレンジング後の全 122,761 件のデータ（旧電カル 73,150 件、新電カル 49,611 件）を利用して意味表現学習を行い、新電カル上位 20 病名のレコードを抽出した後、旧電カル由来のデータを用いて学習モデルを作り、新電カル由来データを評価データとして非線形サポートベクターマシンで評価を行った。マクロ平均 F 値は 0.751 であった。その時の混合行列を表 11.3 に示す。

表 11.3　評価 2 の混合行列

	C15.1	C16.2	C20	C22.0	C25.0	C34.1	C34.3	C34.9	C56	C61	H25.1	H33.0	H35.3	I20.8	I35.0	I48	I50.0	I67.1	M48.06	P03.4
C15.1	457	7	11	2	2	0	0	1	0	0	0	0	0	0	0	0	2	1	0	0
C16.2	13	396	58	15	36	0	0	0	1	3	0	0	0	2	0	6	2	2	0	2
C20	2	1	347	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C22.0	1	3	2	850	5	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0
C25.0	4	0	16	17	464	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
C34.1	2	0	10	0	4	193	71	829	3	4	0	0	0	1	2	4	3	0	0	1
C34.3	0	0	3	1	1	121	65	689	0	4	0	0	0	0	1	2	3	0	2	1
C34.9	1	0	1	0	3	23	24	406	4	0	0	0	0	0	1	3	2	0	0	0
C56	0	0	3	0	0	0	0	0	385	4	0	0	0	0	0	0	1	0	0	0
C61	0	0	1	0	0	0	0	0	0	911	0	0	0	0	0	0	0	0	0	0
H25.1	0	0	0	0	0	0	0	0	0	1	151	74	703	0	0	0	0	0	0	0
H33.0	0	0	0	0	0	0	0	0	0	0	2	298	61	0	0	0	0	0	0	0
H35.3	0	0	0	0	0	0	0	0	0	0	7	28	333	0	0	0	0	0	0	0
I20.8	0	0	1	2	0	9	6	2	2	7	5	1	2	450	78	83	42	6	2	0
I35.0	1	0	0	0	1	1	1	2	1	0	1	0	4	19	506	63	87	1	2	0
I48	0	0	0	0	0	0	0	0	0	0	0	0	1	0	13	453	15	1	0	0
I50.0	0	0	3	0	3	2	3	18	2	5	2	0	1	30	64	160	252	0	0	0
I67.1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	510	1	1
M48.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	372	0
P03.4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	429

最後に、評価 3 として、新電カルデータのみで学習モデルを作成して評価する。データクレンジング後の新電カルデータ 49,611 件に対し意味表現学習を行い、新電カル上位 20 位の診断病名コードを抽出した後、新電カルの評価用データセットを 75 対 25 に分け、75 %を訓練用データセット、残り 25 %を評価用データセットとし、非線形サポートベクターマシンにて評価した。マクロ F 値は 0.874 であった。評価 3 の混合行列を表 11.4 に示す。

評価 1,2,3 の学習モデルによる評価（F 値）の比較を表 11.5 にまとめた。

表 11.4　評価 3 の混合行列

	C15.1	C16.2	C20	C22.0	C25.0	C34.1	C34.3	C34.9	C56	C61	H25.1	H33.0	H35.3	I20.8	I35.0	I48	I50.0	I67.1	M48.06	P03.4
C15.1	119	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C16.2	1	129	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
C20	1	1	85	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C22.0	0	0	0	216	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C25.0	1	2	2	1	120	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C34.1	0	0	0	1	0	215	45	21	0	0	0	0	0	0	0	0	0	0	0	0
C34.3	0	0	0	0	0	66	141	14	0	0	0	0	0	0	1	0	1	0	0	0
C34.9	0	0	1	0	0	40	42	34	0	0	0	0	0	0	0	0	0	0	0	0
C56	0	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	0	0	0
C61	1	0	0	0	0	0	0	0	0	227	0	0	0	0	0	0	0	0	0	0
H25.1	0	0	0	0	0	0	0	0	0	0	208	5	19	0	0	0	0	0	0	0
H33.0	0	0	0	0	0	0	0	0	0	0	17	68	5	0	0	0	0	0	0	0
H35.3	0	0	0	0	0	0	0	0	0	0	44	6	42	0	0	0	0	0	0	0
I20.8	0	0	0	0	0	0	0	0	0	0	0	0	0	152	4	2	17	0	0	0
I35.0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	156	2	10	0	0	0
I48	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	113	6	0	0	0
I50.0	0	0	0	0	0	1	2	2	0	0	0	0	0	13	11	9	98	0	0	0
I67.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	129	0	0
M48.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93	0
P03.4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	107

表 11.5　評価方法の違いによる F 値の比較 (非線形サポートベクターマシンにて評価)

診断病名コード	評価 1 C=100, gamma=0.1	評価 2 C=100, gamma=0.1	評価 3 C=10, gamma=0.1
C15.1	0.707	0.952	0.992
C16.2	0.625	0.913	0.978
C20	0.274	0.911	0.977
C22.0	0.755	0.951	0.995
C25.0	0.142	0.933	0.984
C34.1	0.247	0.255	0.732
C34.3	0.113	0.034	0.696
C34.9	0.335	0.331	0.491
C56	0.863	0.985	1.000
C61	0.949	0.994	0.998
H25.1	0.116	0.468	0.839
H33.0	0.493	0.824	0.809
H35.3	0.333	0.493	0.513
I20.8	0.687	0.831	0.881
I35.0	0.141	0.737	0.908
I48	0.599	0.792	0.921
I50.0	0.547	0.627	0.767
I67.1	0.931	0.994	1.000
M48.06	0.90	0.993	1.000
P03.4	0.961	0.998	1.000
評価件数	11930	11930	2983
マクロ平均 F 値	0.595	0.751	0.874

第 12 章

考察

表 11.5 にて診断病名コード毎の F 値を確認すると、C61、I67.1、M48.06、P03.4 は、評価 1、評価 2、評価 3 のいずれにおいても、0.90 以上の高い F 値を得たが、その他の診断病名コードに関しては評価 1 では C56 が 0.9 に近い F 値を認めたものの全体的に低値である。特に C34.1、C34.3 は 0.2 未満の低値であり、表 11.2 の混合行列で予測値の確認を行った。すると C34.1、C34.3 のいずれも予測値が C34.9 に割り振られている事が判明した。新電カルデータで学習モデルを作成する評価 3 では、C34.1 の F 値が 0.732、C34.3 の F 値が 0.696 に改善していることから、旧電カルでは、詳細不明コードである「.9（ドットナイン)」を多用していたが、新電カルでは詳細分類までコード付けを行うようになったことが旧電カル学習モデルにおいて F 値が低値であることの要因であったと推察された。

DPC/PDPS 傷病名コーディングテキストにおいて詳細不明コード「.9」の付与は例外的な事例とされており、適切な診療録の記載や主治医へ確認するなど適切な確認体制を構築することが求められている [41)]。新電カルでは詳細不明コードを選択すると、詳細分類コードの入力を促すようにシステムを変更した。その効果があったのか新電子カルで学習モデルを作成する評価 3 では、詳細分類コードまで学習されマクロ平均 F 値 0.874 に改善した。CAC 構築のためにはコーディングの入力ルールを理解し、詳細分類コードまで入力することが重要であると思われた。本研究によって CAC 学習モデルを構築する際に、「.9」の付与が予測結果に大きな影響を与える結果となった。

次に、眼及び付属器の疾患に属する ICD10 コードに着目する。表 8.2 を確認すると、新電カルデータセットにおいて H40.1（原発開放隅角緑内障）の登録件数が 798 件で登録件数が 7 位であったものが、表 10.1 のデータクレンジング後の新電カルデータセットでは、上位 20 病名以内にランキングされていないことが判明した。データセットを確認すると、眼科の経過要約が 50 文字に満たないレコードが多数確認でき、データクレンジング時に除外していたことが原因であった。眼科の特殊性として短期滞在手術による 1 泊 2 日の入院の場合に短い経過要約となる場合があり、眼科の退院サマリーを処理する際には、経過要約が短い傾向があることに留意する必要があろう。

表 10.3 において H25.1（老人性核白内障）の旧電カルの登録件数が 123 件と学習用の

データが少ないながら、表 11.2 の混合行列を確認すると、多くの場合 H35.3（黄斑及び後極の変性）を予測していたことがわかる。H33.0（網膜剥離、網膜裂孔を伴うもの）、H35.3（黄斑及び後極の変性）を見ても多くの場合は H35.3（黄斑及び後極の変性）を予測している。概ね「第 VII 章　眼及び付属器の疾患（H00-H59）」への振り分けがされていることが確認できたが、評価 1 では眼科の詳細分類コードまで予測することが困難であった。評価 1 と新電カルのみのデータで学習モデルを構築する評価 3 を比較すると、比較表（表 11.5）より H25.1 が 0.116 から 0.839、H33.0 が 0.493 から 0.809、H35.3 が 0.333 から 0.513 といずれも改善されている。このように新電カルの導入移行は、診断病名の詳細分類においても学習が進み F 値の向上が確認できた。表 10.1 より H25.1 は新電カルにおいてデータ数が 929 件と十分にあることから、F 値が一段と向上したのではと考える。各詳細分類コードまで含む退院サマリーのレコード数と F 値との関連性は今後引き続き調査していきたい。

評価 1 ではマクロ平均 F 値 0.595 であったものが、評価 3 ではマクロ平均 F 値が 0.874 であった。表 10.3 に示すように新旧電カルでモデルの分布が大きく異なるにも関わらず、0.6 程度のマクロ平均 F 値を得たことは異なる電子カルテシステム間でも汎用的に CAC が動作していたことと解釈でき、CAC の構築に意味表現学習を採用する選択理由がより高まったといえる。

その他、旧電カルと新電カルでは退院サマリーの入力画面が大幅に異なっている（図 8.1, 図 8.2）。共通の記載項目が包含されていても、入力画面が異なることで経過要約の内容が変化した可能性は否定できない。本研究の限界として、扱った電子カルテシステムが 2 種類のみであるため検証はできないが、今後、対象を他社の電子カルテシステムで入力されたデータにも対象範囲を拡大し検証していくことも視野にいれたい。退院サマリー作成に関するガイダンスではプロファイルからの引用項目等まで言及されており、各社の電子カルテシステムにこのガイダンスで示された仕様が浸透していくことで CAC の性能検証としてもスムーズに進むことが期待された。

CAC 構築手法の考察を行う。我々は富山大学附属病院の 16 年分の経過要約に医療辞書導入による形態素解析及び意味表現学習を施し、264 種類の特徴単語ベクトル値を求めた。そこに標準退院サマリーの項目である「性別」、「退院時年齢」、「診断病名」、「経過要約」、及び記載者情報と関連性の強い項目として「診療科名」を説明変数として追加し、目的変数を診断病名コードとし、サポートベクターマシンによる機械学習を行った。本研究のように意味表現学習を用いて特徴単語ベクトル値を説明変数として扱い学習モデルを構築した事例は、医療分野では過去に例がない。評価 3 では同一のシステムでの予測モデルを作成した。意味表現学習を利用した新電カル学習モデルで評価し、主要診断病名コード上位 20 病名のマクロ平均 F 値が 0.874 であった。評価 2 は、新旧の電カル

データを結合したもので意味表現学習までを行ったもので、マクロ F 値が 0.751 と、評価 1 と評価 3 の中間のマクロ平均 F 値であった。

また、注目されている BERT を用いた報告として荒牧 [42] らが作成した医療文書の大規模コーパスを用い、柴田 [43] らの研究グループはその大規模な医療文書に BERT を用い文脈理解の精度検証事例を報告している。東京大学医学部附属病院の 1 億 2 千万もの医療文書（UTH-BERT）に対して、「疼痛」という表現に対して事象認識と事実性判定を行う抽出器を構築し、精度を評価した。疼痛が現在問題となっている疼痛であるかどうか、コントロールされている疼痛なのか、また事実によって記載されたものなのか、推論によって判断されるものなのかなど、疼痛の事実を区別するタスクを定義して、BERT を用いて精度を評価した。結果、医療文書で事前学習を行った BERT を用いることで、事象認識と事実性判定を伴う疼痛表現の抽出精度が向上することを示した。

今回の柴田らの研究は CAC の構築を目的としたものではないにせよ、医療文書の文脈判定という観点から、退院サマリーに対し BERT を適応させれば、CAC への応用発展の可能性がある。BERT には入力長が 512 トークンまでという限界がある。図 8.4 をみると電子カルテシステムに応じて退院サマリー経過要約への入力文字数に違いが認められそうだが、経過記録が SOAP 形式で記載される以上、512 トークンもあれば判定には十分ではないだろうか。BERT は 2018 年公開の新しい手法であり、大学院に在学期間中は BERT での実験ができなかったが、どの手法が CAC として汎用的に動作するかは今後、同条件下での各手法の比較研究をしていきたい。

意味表現学習は 264 種類の特徴単語と経過要約の関連性が数値化できる。田中 [44] は具体的な診療記録の例を挙げ 264 種類の特徴単語の中で医療に関係の深い特徴単語を抽出し、経過要約と特徴単語との関連性を数値化して提示した。我々の提案する意味表現学習とサポートベクターマシンの組み合わせによる CAC 構築手法は、導いた予測値を人間が理解しやすい形で数値表現できることが特徴であり、説明責任を重要視する医療分野において親和性のある手法と考えた。

次に今後の拡張の可能性を考察する。例えば診療情報管理士の質的点検業務に応用できるのではないかと考える。質的点検業務では診療科毎に退院サマリーの内容が適正に記載されているかのチェックを行う。そこで正しく記載されているか否かを退院サマリーに正解ラベルを付与することで、機械学習による、正誤予測が可能になるだろう。またラベルを付与しない状態の教師なし機械学習の場合でも、クラスター分析を行うことで、正しく記載されている群と記載されていない群が可視化できる可能性がある。現在、診療情報管理士はすべての退院サマリーを目視確認し、退院サマリが正しく記載されているかを目視にて確認しているが、必ずしも全件をチェックする必要がなくなり、診療情報管理士の業務改善につなげることができるのではなかろうか。

最後になるが、本研究を遂行しながら、改めて医療制度の変遷が激しい事を痛感した。この 20 年間で DPC/PDPS の導入、それに伴う ICD10 コーディングの高い精度を求められ、さらに、退院サマリーの退院後 2 週間以内に作成するという迅速な記載の要求が起因となり、医師がより多忙化、結果、サマリに記載する文字数も大幅に減ったことが、データ分析やデータクレンジング時に眼科カルテが削除されてしまった所などからも伺えた。診療情報管理士はこのような医療制度の変更には常にアンテナが高い状態である。その意味でも、診療情報管理士自身が機械学習のモデルを作成することができるようになり、自ら予測モデルの評価ができるようになってほしいと考えた。理由は現時点で CAC として良いものが出来たとしても、数年後には制度の変更から使い物にならない場合も想定される。その時に、診療情報管理士、自ら原因を探求し、時代にそった CAC に適時アップデートできる職種になって頂きたいと願う。

筆者は、日本病院会認定診療情報管理指導者として、診療情報管理士の教育に携わっているが、最新の技術を習得するにとどまらず、社会情勢などバランス良く知識を反映していくことが大切なのだと実践を通じて認識できたのは大きな収穫であった。是非、福井工業大学大学院で得た経験を糧にし、今後の教育活動に活かしていきたいと考えた。

第 13 章

まとめ

経過要約を意味表現学習することで得られた「264 個の特徴単語ベクトル値」と標準退院サマリーの入力項目でもある「年齢」「性別」「診療科名」、「診断病名コード」を目的変数とした、非線形サポートベクターマシンによる本邦版 CAC の構築手法を提案した。

本研究の貢献は、意味表現学習を採用し、医療文書をニューラルネットワークにより特徴単語毎のベクトル値に変換することで、結果を人間が理解しやすい形で表現することが可能となり、判定結果の解釈性の向上が期待できる構築手法を提案したこと。また、学習時の説明変数の設定を可能にし、混合行列の結果より診療情報管理士の知見を活かした評価を可能にした点である。さらに 2 種類の電子カルテシステムの退院サマリーデータを用い機械学習モデルの汎用的な評価が行えるようなベンチマークの手法を確立した。

新電カルの記載上位 20 病名に対し評価した。旧新電カルでそれぞれ意味表現学習にて学習モデルを構築した場合マクロ F 値が 0.595 であった。旧新電カルを合わせて意味表現学習にて学習モデルを構築した場合、マクロ平均 F 値 0.751 であった。新電カルのみのデータを用いて意味表現学習にて学習モデルを構築した場合マクロ F 値が 0.874 であった。　結果の差異の要因として両者のモデルの分布が異なる事が考えられた。詳細不明コード「.9」が付与された状態で学習モデルを作成すると、詳細分類毎の F 値に負の影響を与えたが詳細分類コードまで入力することで、F 値の改善が示唆された。

本研究は、JSPS 科研費 JP20K11833 の助成を受け実施した。また富山大学附属病院臨床研究倫理審査委員会の承認（R2018004）及び福井工業大学研究倫理審査委員会の承認（人-2019-06）を受け実施した。利益相反はない。

謝辞

本研究を遂行するにあたり、様々な方のご協力を賜りました。博士後期課程入学以前から数年間にわたり、事前準備を含め実験から評価、さらには論文投稿や、研究の遂行を指導して頂いた芥子育雄教授をはじめ、同じく芥子研究室 OB のオオイ・コックリオンさんと田中佑樹さんには、様々な辞書を利用した事前学習の結果等の情報提供を頂き、その後の実験をスムーズ進めることができました。感謝の気持ちでいっぱいです。

社会福祉士の国家試験や診療情報管理士の認定試験等、様々な準備に追われる中、卒業研究のテーマとして形態素解析用の医療辞書の作成に取り組んでくれた、金城大学辻岡ゼミＯＧの飴井早紀さん、五十嵐優奈さん、小林愛未さん、白浜夏紀さん達とは共に苦労を分かち合いました。

富山大学の中川肇名誉教授、林篤志附属病院長には激励の言葉を沢山頂いただけでなく共著者としてもご助言を頂戴致しました。医療情報学会中部支部会の浜松医科大学の木村通男教授、愛知淑徳大学の加藤憲教授、福井大学医学部附属病院の山下芳範准教授、関係者の皆さまには、年度末のお忙しい中、口頭発表の機会を与えていただきました[45)]。

本年度より赴任した国立国際医療研究センターにおいては、美代賢吾センター長を初め、既に多くの先生方から大変なお力添えを賜り、様々な研究に参画させて頂いているところです。その中で特にゲノム分析事業等には本研究で得た知見を参画中の研究に還元できる部分もありそうで、今後も深く研究を進めていきたいと考えています。その他多くの方から応援して頂きました。この場をかりて感謝申し上げます。

参考文献

1) 辻岡和孝, 芥子育雄, 中川肇, 林篤志. 自然言語処理を利用した本邦版 ComputerAssistedCoding 構築手法の研究, 診療情報管理 Vol34(1),pp56-64,2022

2) Ryan Sandefer et al. Survey Predicts Future HIM Workforce Shifts: HIM Industry Estimates the Job Roles, Skills Needed in the Near Future,Journal of AHIMA vol.86(7) pp32-35.

3) 3M:Improving Healthcare Data Quality Using Computer-Assisted Coding(2019), https://multimedia.3m.com/mws/media/1690506O/cac-brochure.PDF (cited 2021-April-1)

4) Sharon Campbell et al. A narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals, Health Information Management Journal vol49(1), pp5-18,2020

5) Kathleen C. Fraser et al. Extracting umls concepts from medical text using general and domain-specific deep learning models., In Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis,pp157–167,2019.

6) 鈴木隆弘, 土井俊祐, 他. 多施設テキストデータベースを用いた退院時サマリー作成・監査支援の試み, 医療情報学 Vol.39,pp667-668,2019

7) 鈴木隆弘, 土井俊祐, 他. 退院サマリー監査を支援する DPC 判定アプリケーション, 医療情報学 Vol.38,pp786-787,2018

8) 野口怜, 鳥飼幸太, 齋藤勇一郎. 新しい自然言語処理エンジンを用いた電子カルテデータ構造化と疾患別特徴語抽出, 日本医療情報学会春季学術大会シンポジウム 2019 in 熊本抄録集 vol.39,pp64-65,2019.

9) 木村 知広, 津本 周作, 平野 章二. 機械学習による退院時要約からの DPC 分類の推測, 医療情報学 vol.40,pp700-704,2020.

10) HL7CDA に基づく退院時サマリー規約, http://helics.umin.ac.jp/helicsStdList.html(2019/10/16 認定) (cited 2021-April-1)

11) 退院サマリー作成に関するガイダンス, (2019 年 9 月), http://jami.jp/jamistd/docs/dischargeSummary2019.pdf (cited 2021-September-1)

12) 一般社団法人 MedicalExcellenceJAPAN. 国民のための合理的医療を追求するツールとしての電子カルテシステムの改革にむけた提言　世界をリードするデジタル医療基盤を目指して, 四次元医療改革研究会電子カルテ分科会提言,2021

13) 石川ベンジャミン光一, 中田悠太, 辻岡和孝, 堤ともゑ. 次世代の診療情報管理士から-データ分析を中心に-, 診療情報管理 Vol34(3),pp43-62,2022
14) 渡辺直．電子カルテ時代の POS ―患者指向の連携医療を推進するために―. 医学書院,pp150-154，2012
15) 辻岡和孝, 中川肇. 富山大学附属病院における電子クリニカルパスの導入経験-稼働 1 年後の評価-, 富山大学医学会誌 Vol.26(1),pp667-668,2015
16) 厚生労働省大臣官房統計情報部,ICD の ABC, 厚生労働省, https://www.mhlw.go.jp/toukei/sippei/dl/icdabc_r04.pdf(cited 2022-April-1)
17) 辻岡和孝, 渡邊翔太, 片口治幸, 中川肇. パスの評価内容をサマリに自動展開する機能の提案, 日本クリニカルパス学会学会誌抄録集，19th，p.408，函館市，2018
18) 辻岡和孝, 鍋島一斗, 中川肇. 退院調整と電子パス利用率との関連性の調査報告, 日本クリニカルパス学会学会誌抄録集，18th，p.479，大阪市，2017
19) Devlin J, Chang M, Lee K, Toutanova K. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding,arXiv: 1810.04805. 2018. Available from: https://arxiv.org/abs/1810.04805 (cited 2021 May 31)].
20) Qiita.SVM を勉強してみる, https://qiita.com/shuva/items/7ded094257417677d51f (cited 2022-April-1)
21) 大江和彦, 今井健. 臨床医学知識処理を目指した医療オントロジー開発, 人工知能学会誌 Vol25(4),pp493-500,2010
22) 大江和彦. 臨床医学オントロジーの応用と今後の展開, https://www.m.u-tokyo.ac.jp/medinfo/medont2009projA/s20100330-7.pdf (cited 2022-April-1)
23) 吉崎亮介.【キカガク流】人工知能・機械学習脱ブラックボックス講座（中級編）,Udemy, 2021.
24) 中山光樹. 機械学習・深層学習による自然言語処理入門, マイナビ出版, 2020.
25) 松岡亮二. 不平等な学歴獲得競争,WASEDA ONLINE, https://yab.yomiuri.co.jp/adv/wol/opinion/society_160411.html(cited 2022-April-1)
26) Taku Kudo. MeCab: Yet Another Part-of-Speech and Morphological Analyzer,http://taku910.github.io/mecab/ (cited 2021-April-1)
27) 学研究費助成金臨床研究等 ICT 基盤研究事業 MEDNLP ホームページ：http://sociocom.jp/data/2018-manbyo/index.html(cited 2021-April-1)
28) Comejisyo プロジェクトホームページ：https://osdn.jp/projects/comedic(cited 2021-April-1)
29) 辻岡和孝, 渡邊翔太, 片口治幸, 中川肇. ニューラルネットワークを用いた単語ベクト

ルの演算による ICD10 の抽出, 診療情報管理 Vol30(4),pp70-72,2019

30) Mikolov T, Chen K, Corrado G and Dean J.Efficient estimation of word representations in vector space, International Conference on Learning Representations 2013 Workshop Proceedings , 2013.

31) 坪井祐太, 海野裕也, 鈴木潤. 深層学習による自然言語処理, 第 1 版, 講談社, 東京,2017

32) 田中昌昭. 単語の分散表現を用いた文書分類, 川崎医療福祉学会誌,vol.28(1),p167-178,2018

33) 後藤貴樹.BERT を用いたクラスタ分析による文書分類, 岐阜工業高等専門学校電気情報工学科卒業研究報告,2022(cited 2022-June-20)

34) 堂坂浩二，石井雅樹，伊東嗣功. 深層学習による医療文書からの病名と医療行為の抽出, 秋田県立大学ウェブジャーナル vol.6,pp209-215,2019.

35) Y. Sakaizawa and M. Komachi. Construction of a Japanese Word Similarity Dataset, CoRR, abs/1703.05916, 2017

36) I. Keshi, Y. Suzuki, K. Yoshino, S. Nakamura. Semantically Readable Distributed Representation Learning and Its Expandability Using a Word Semantic Vector Dictionary, IEICE Trans. on Information and Systems, Vol.E101-D, No.4, pp.1066-1078, April 2018.

37) 芥子育雄, 鈴木優, 吉野幸一郎, 大原一人, 向井理朗, 中村哲. 分散的意味表現学習のための単語意味ベクトル辞書 Ver.2 と日本語 Twitter 極性分析ベンチマークについて, 情報処理学会研究報告, Vol. 2017-NL-231, Vol. 2017-SLP-116, No. 8, pp.1-7, 2017.

38) 芥子育雄，松田義貴，鈴木優，吉野幸一郎，中村哲．意味表現学習におけるツイート分散表現の解釈性評価と可視化の提案，第 10 回データ工学と情報マネジメントに関するフォーラム, DEIM Forum 2018, pp.B1-1,2018.

39) M. Faruqui, J. Dodge, S.K Jauhar, C. Dyer, E. Hovy, N. A. Smith. Retroftting Word Vectors to Semantic Lexi-cons, In Proc. of NAACL-HLT, pp. 1606-1615, 2015.

40) オオイ　コックリオン. 意味表現学習による解釈性のある病名推定に関する研究, 福井工業大学電気電子工学科卒業研究発表会予稿集,pp49-50,2020

41) 厚生労働省.DPC ／ PDPS 傷病名コーディングテキスト, 厚生労働省ホームページ:https://www.mhlw.go.jp/content/12404000/000668757.pdf(cited 2021-April-1)

42) 荒巻英治, 若宮翔子, 矢野憲, 他. 病名アノテーションが付与された医療テキスト・コーパスの構築. 自然言語処理 Vol.25(1),pp119-152,2018.

43) 柴田大作, 河添悦昌, 嶋本公徳, 篠原恵美子, 荒牧英治. 診療記録で事前学習した BERT による疼痛表現の抽出病名. 医療情報学 Vol.40(2),pp73-82,2020.

44) 田中佑樹. 新しく導入した医療辞書を使った病名予測に関する研究, 福井工業大学電

気電子工学科卒業論文,2021

45) 辻岡和孝, 芥子育雄, 中川肇, 林篤志. 意味表現学習とサポートベクタマシンを用いたCAC システムの評価, 第 17 回日本医療情報学会中部支部学術集会抄録集,p4,2022

研究業績リスト

主論文に関する研究業績

【　原　著　】

[1] 辻岡和孝, 芥子育雄, 中川肇, 林篤志. 自然言語処理を利用した本邦版 ComputerAssistedCoding 構築手法の研究, 診療情報管理，vol.34(1)，pp.56-64，2022. 査読有

[2] 辻岡和孝, 中川肇. 富山大学附属病院における電子クリニカルパスの導入経験―稼働１年後の評価―, 富山大学医学会誌，vol.26，pp.33-38，2015. 査読有

【　研究速報　】

[1] 辻岡和孝，渡邊翔太，片口治幸，中川肇. ニューラルネットワークを用いた単語ベクトルの演算によるＩＣＤ１０の抽出, 診療情報管理，vol.30(4)，pp.70-72，2019. 査読有

【　学会発表　】

[1] ○辻岡和孝，芥子育雄，中川肇，林篤志．一般演題「意味表現学習とサポートベクターマシンを用いた CAC システムの評価」，第 17 回日本医療情報学会中部支部会，2022 年 3 月 26 日　，オンライン開催

[2] ○石川ベンジャミン光一，○中田悠太，○辻岡和孝，○堤ともゑ．シンポジウム　次世代の診療情報管理士から　－データ分析を中心に－「ＡＩ・ＲＰＡ時代の診療情報管理業務を見据えて」，第 47 回日本診療情報管理学会学術大会，2021 年 9 月 17 日，オンライン開催

[3] ○辻岡和孝．パネルディスカッション「退院時サマリーを機械学習することで得られる病名推定技術の中間報告」，第 52 回北陸診療情報管理研究会，2019 年 11 月 16 日，能美市

その他の業績

【　論　文　】

[1] 中川肇,辻岡和孝. 地域医療連携医開業医の診療情報の参照場面と種別の調査報告-1 年間のアクセスログの分析から開業医の参照のニーズを考察する, 診療情報管理，vol.31(1)，pp.61-64，2019. 査読有

[2] 中川肇, 古澤桂子, 渡邊翔太, 片口治幸,辻岡和孝. 医療経済的にみた本院における患者のキャンセル率と予防の一方策, 富山大学医学会誌，vol.29，pp.12-16，2019. 査読有

[3] 辻岡和孝, 中川肇.DPC 公開データと看護必要度データを活用した MDC・ポートフォリオ・マネジメント , 診療情報管理，vol.29(4)，pp.40-44，2018. 査読有

[4] 中川肇,辻岡和孝. 電子化文書管理における長期署名フォーマットの変更時の証跡性担保に関する方策, 医療情報学，vol.37(4)，pp.179-185，2017. 査読有

[5] 辻岡和孝, 中川肇. 携帯情報端末を用いた電子カルテへの写真登録システムの構築と評価, 医療情報学，vol.37(4)，pp.197-204，2017. 査読有

[6] 辻岡和孝.HIS リプレイス作業でのシステム拡張性を検討する, 月刊新医療，9 月号，pp.62-64，2009. 査読無

[7] 辻岡和孝. 医療機器管理システムの構築とその意義, 月刊新医療，2 月号，pp.109-111，2006. 査読無

【　学会発表　】

[1] ○渡邊佳代, ○宮沢春菜, ○香川璃奈, ○辻岡和孝, ○大竹正規．ワークショップ　運動・口腔保健・栄養・休養・自己管理のための保健医療情報，第 41 回医療情報学連合大会，2021，名古屋市

[2] ○辻岡和孝, 中川肇．一般口演　電子カルテの真正性向上を目的としたブロックチェーン技術の応用 ，第 14 回日本医療情報学会中部支部学術集会，2019，金沢市

[3] ○辻岡和孝, 渡邉翔太, 片口治幸, 中川肇．ポスター　動画投稿サイト vimeo を利用した病院職員を対象としたセキュリティ教育の取組み ，第 38 回日本医療情報学連合大会，2018，福岡市

[4] ○辻岡和孝, 渡邉翔太, 片口治幸, 中川肇．ポスター　パスの評価内容をサマリに自動展開する機能の提案，第 19 回日本クリニカルパス学会，2018，金沢市

[5] ○辻岡和孝, 渡邉翔太, 片口治幸, 中川肇．一般口演　電子カルテ稼働 10 年間におけるテンプレート作成及び利用状況 ，第 44 回日本診療情報管理学会学術集会，2018，新潟市

[6] ○辻岡和孝, 渡邉翔太, 片口治幸, 中川肇．一般口演　過去 10 年間の正規化された退院サマリを機械学習することで得られた病名推測システムの構築 ，第 44 回日本診療情報管理学会学術集会，2018，新潟市

[7] ○辻岡和孝, 渡邉翔太, 片口治幸, 中川肇．一般口演　Moodle を利用した医療情報教育システムの構築，第 13 回日本医療情報学会中部支部学術集会，2018，浜松市

[8] ○辻岡和孝, 鍋島一斗, 中川肇．ポスター　入院期間を意識したパス運用におけるＰＤＣＡサイクルの必要性，平成 29 年度大学病院情報マネジメント部門連絡会議，2018，旭川市

[9] ○辻岡和孝, 鍋島一斗, 中川肇．ポスター　OpenSourceSoftware とフリーウェアを活用した部門システム向けセキュリティ設計の一提案 ，第 37 回医療情報学会連合大会，2017，大阪市

[10] ○辻岡和孝, 中川肇, 安藤基也, 後藤秀紀, 音川奈雄也, 上田理絵, 柳浦一博．ポスター　病院情報システムの利用者心得（HI-UP）の医師事務作業補助者研修への応用 ，第 37 回医療情報学会連合大会，2017，大阪市

[11] ○辻岡和孝, 中川肇, 安藤基也, 後藤秀紀, 音川奈雄也, 上田理絵, 柳浦一博．ポスター　退院調整と電子パス利用率との関連性の調査報告，第 18 回日本クリニカルパス学会，2017，大阪市

[12] ○辻岡和孝, 中川肇．一般口演　積雪が外来患者数に及ぼす影響についての調査報告，第 43 回日本診療情報管理学会学術集会，2017，札幌市

[13] ○薄井勲, 小清水由紀子, 朴木久恵, 藤坂志帆, 瀧川章子, 岡部圭介, 角朝信, 岩田実, 石木学, 安井真希, 鍋山昭子, 坂本純子, 角田美鈴, 荒俣文恵, 今井あゆみ, 高邑小百合, 辻岡和孝, 中川肇, 戸邉一之．ポスター　既製品のデータ解析ソフトを利用した電子カルテへの SMBG の簡易取り込みシステムの構築，第 60 回日本糖尿病学会年次学術集会，2017，名古屋市

[14] ○辻岡和孝, 中川肇．一般口演　携帯端末のカメラ機能を用いた電子カルテ画像登録システムの評価，第 36 回医療情報学会連合大会，2016，横浜市

[15] ○辻岡和孝, 中川肇．ポスター　電子パス 1 年後のアウトカム入力率の調査報告，第 17 回日本クリニカルパス学会学術集会，2016，金沢市

[16] ○中川肇,辻岡和孝, 鍋島一斗, 北本史穂, 穴田博史, 後藤秀樹．一般口演　電子カルテのリプレースに関する知見-データ移行とリハーサル (操作訓練)-，第 10 回日本医療情報学会中部支部学術集会，2016，金沢市

[17] ○中川肇,辻岡和孝, 村井貴子, 齊藤夕佳乃．一般口演　長期署名フォーマット PAdES から XAdES への移行と業務拡大の経験，第 35 回日本医療情報学連合大会，2015，宜野湾市

[18] ○辻岡和孝, 中川肇, 薄井勲, 戸邉一之．一般口演　NFC 血糖測定器を用いた，電子カルテへの転記と指導管理料対策システムの構築，第 35 回日本医療情報学連合大会，2015，宜野湾市

[19] ○鵜野浩靖, 中川肇, 辻岡和孝, 瀬戸美和子, 山田真治, 堂本照貴, 梨木康平, 花井剛紀, 森淳．一般口演　医療機関における院内無線 LAN によるインターネット接続環境の提供，平成 26 年度大学病院情報マネジメント部門連絡会議，2015，岐阜市

[20] ○高木英子，中山眞由美，辻岡和孝．ポスター　電子カルテフェリカポートの活用　フェリカ測定器によるバイタルサインの入力，平成 25 年度大学病院情報マネジメント部門連絡会議，2014，徳島市

[21] ○辻岡和孝, 中川肇, 村井貴子, 齊藤夕佳乃, 魚住恭子, 平野千春．ポスター　退院時サマリー承認率向上のためのサポート体制の強化とその評価，平成 25 年度大学病院情報マネジメント部門連絡会議，2014，徳島市

[22] ○辻岡和孝, 中川肇．一般口演　無線通信機能付医療機器の運用方法の検討，第 8 回日本医療情報学会中部支部学術集会，2013，津市

[23] ○辻岡和孝．一般口演　医療機器管理システムの構築とその意義，第 33 回臨床工学研修会，2008，富山市

[24] ○辻岡和孝，齋藤哲哉，塗茂裕一，大和田利郎，深沢卓，加藤英治．一般口演　電子カルテリプレース作業におけるシステム拡張性の検討，第 28 回医療情報学会連合大会，2008，横浜市

[25] ○辻岡和孝，五十嵐茂幸，土屋良武．一般口演　クオリティマネジメントシステムにおける医療機器管理システムの位置付け，第 25 回医療情報学会連合大会 ，2005，横浜市
[26] ○辻岡和孝，五十嵐茂幸，出口繁雄，吉村美香，米田陽子，土屋良武．一般口演　統合型医療機器管理システムの開発，第 15 回日本臨床工学会，2005，札幌市
[27] 佐藤和孝，市原清志，○辻岡和孝，西井研治．ポスター　健康危険度調査票を利用した実践的な健康教育支援システムの開発，第 18 回医療情報学会連合大会 ，1998，神戸市

２６４種類の特徴単語

0，人間，human 1，人名，person_name 2，男性，man 3，女性，woman 4，子供，child 5，大人，adult 6，老人，elderly 7，家族・家庭，family 8，動物，animal 9，水棲生物, aquatic_organism 10，鳥類，bird 11，虫，insect 12，微生物，microbe 13，植物，plant 14，生命，life 15，生死，life_and_death 16，誕生，birth 17，病気，disease 18，老い，old_age 19，殺生，killing 20，性，gender 21，人間の身体, human_body 22，内臓器官, visceral_organ 23，生物の身体, creature_body 24，娯楽・趣味，hobby 25，スポーツ，sport 26，旅行，travel 27，活動，activity 28，共同，collaboration 29，食，eclipse 30，住居，dwelling 31，衣類，clothes 32，健康・美容，health 33，日用品，daily_necessities 34，装飾品，decoration 35，道具，tool 36，機械・機器，machine 37，建造物，structure 38，通信，communication 39，マスメディア，mass_media 40，交通・輸送，traffic 41，自動車，car 42，船舶，ship 43，航空機，aircraft 44，会社・職業，company 45，教育・育児，education 46，福祉・年金，welfare 47，施設・設備，facility 48，公共制度, public_system 49，法律・法令，law 50，税制，tax_system 51，社会活動，social_activity 52，流行・人気，fashion 53，社会問題，social_problem 54，犯罪，crime 55，性問題，sex_matter 56，国家，country 57，政治，politics 58，政府・省庁，government 59，財政，finance 60，経済，economy 61，金融，financing 62，エネルギー問題，energy_problems 63，外交，diplomacy 64，軍事・防衛，military 65，平和，peace 66，戦争・紛争，war 67，国際関係，international_relations 68，地名，name_of_place 69，国名，country_name 70，日本，japan 71，都会，city 72，地方，region 73，海外，overseas 74，アジア，asia 75，中近東，near_and_middle_east 76，欧州，europe 77，アフリカ，africa 78，北米，north_america 79，中南米，central_and_south_america 80，オセアニア，oceania 81，極地，farthest_land 82，陸地，land 83，山岳地，mountain_land 84，天空，air 85，海洋，ocean 86，気象・気候，weather 87，環境，environment 88，災害，disaster 89，他の自然，nature 90，地球，earth 91，天体，heavenly_body 92，感覚，sense 93，感情，emotion 94，喜楽，enjoyment 95，悲哀，grief 96，恐怖，fear 97，変化，change_for_sql 98，秩序・順序，order_for_sql 99，数量，quantity 100，勢力・程度，power 101，価値・質，worth 102，因果，causal 103，優良，excellent 104，劣悪，coarseness 105，肯定的，positive 106，否定的，negative 107，新しさ，newness 108，古さ，oldness 109，多数・多量，majority 110，小数・少量，decimal_fraction 111，美麗，beauty 112，醜悪，ugliness 113，一般・平凡，general 114，特殊・希有，special 115，高速，speedy 116，低速，low_gear 117，単純，simple 118，複雑，complicated 119，容易，easy 120，困難，difficult 121，安価，cheap 122，高価，high_price 123，強力，strong 124，脆弱，fragile 125，正確，accurate 126，誤謬，mistake 127，理性的，rational 128，感情的，emotional 129，聡明，intelligent 130，優しさ，kindness 131，組織，organization 132，個人，individual 133，人工，artificial 134，天然，natural_for_sql 135，実存，existence 136，実質・本質，essence 137，性質，property 138，所有，possess 139，動作，

action 140 , 行為 , behavior 141 , 運動 , movement 142 , 停止 , stop 143 , 動的 , dynamic 144 , 静的 , static 145 , 蒸発・気化 , evaporation 146 , 凝固・凍結 , freezing 147 , 溶解・液化 , melting 148 , 発熱・発光 , fever 149 , 燃焼 , burning 150 , 反応 , reaction 151 , 他の現象 , phenomena 152 , 鉱物 , mineral 153 , 資源 , resource 154 , 素材・材料 , material 155 , 重量・質量 , mass 156 , 色彩 , color 157 , 固体 , solid 158 , 気体 , gas 159 , 液体 , liquid 160 , 重さ , weight 161 , 軽さ , lightness 162 , 堅固 , substantial 163 , 柔軟 , flexible 164 , 熱暑 , hot 165 , 温かさ , warmness 166 , 寒冷 , coldness 167 , 明るさ , brightness 168 , 暗さ , darkness 169 , 白 , white 170 , 黒 , black 171 , 赤 , red 172 , 青 , blue 173 , 黄 , yellow 174 , 緑 , green 175 , 無色・透明 , transparent 176 , 空間 , space 177 , 平面 , plane 178 , 立体 , three_dimensional 179 , 球体 , sphere 180 , 形状 , shape 181 , 大規模 , large_scale 182 , 小規模 , small_scale 183 , 長さ , length 184 , 短さ , brevity 185 , 高さ , height 186 , 低さ , low 187 , 厚さ , width 188 , 薄さ , slenderness 189 , 広大 , vast 190 , 狭窄 , contraction 191 , 鋭さ , sharpness 192 , 鈍さ , dullness 193 , 細密 , detailed 194 , 粗雑 , crude 195 , 細さ , thinness 196 , 太さ , thickness 197 , 時間・年月 , time 198 , 季節 , season 199 , 現在 , present 200 , 過去 , past 201 , 未来 , future 202 , 民族・人種 , race 203 , 知識 , knowledge 204 , 言論・発話 , discussion 205 , 書物・書籍 , book 206 , 思想・哲学 , idea 207 , 倫理・道徳 , ethics 208 , 宗教 , religion 209 , 考古学 , archaeology 210 , 歴史 , history 211 , 文学 , literature 212 , 人類学 , anthropology 213 , 心理学 , psychology 214 , 地理 , geography 215 , 言語 , language 216 , 風俗・習慣 , custom 217 , 文化 , culture 218 , 伝統 , tradition 219 , 芸術 , art 220 , 映像・画像 , image 221 , 音楽 , music 222 , 数学 , mathematics 223 , 物理学 , physics 224 , 天文学 , astronomy 225 , 地学 , earth_sciences 226 , 医学・薬学 , medicine 227 , 生物学 , biology 228 , 製造・工作 , manufacture 229 , 機械工学 , mechanical_engineering 230 , 土木・建築 , civil_engineering 231 , 開発 , development 232 , 電気工学 , electrical_engineering 233 , 電子工学 , electronics 234 , コンピュータ , computer 235 , 半導体 , semiconductor 236 , ハード , hardware 237 , ソフト , software 238 , システム , system 239 , oa , oa 240 , 音響 , sound 241 , 通信技術 , communication_technology 242 , 生物・バイオ , bio 243 , 化学 , chemistry 244 , エネルギー , energy 245 , 石油・鉱物 , petroleum 246 , 原子力 , nuclear_power 247 , 軍事技術 , military_technology 248 , 航空・宇宙 , aerospace 249 , 農業 , agriculture 250 , 林業 , forestry 251 , 水産業 , fisheries 252 , 工業 , industry 253 , 鉱業 , mining 254 , 生産 , production 255 , 計画 , plan 256 , 製造業 , manufacturing_industry 257 , 建設業 , construction_industry 258 , 商業・貿易 , commerce 259 , サービス業 , service_industry 260 , 設計・デザイン , design 261 , 宣伝広告 , advertisement 262 , 顧客・ユーザ , customer 263 , 出版業 , publishing_xbusiness