Natural Language Processing Series

自然言語処理シリーズ 12

医療言語処理

工学博士 奥村 学 監修

博士(情報理工学) 荒牧 英治 著

コロナ社

刊行のことば

人間の思考，コミュニケーションにおいて不可欠なものである言語を計算機上で扱う自然言語処理という研究分野は，すでに半世紀の歴史を経るに至り，技術的にはかなり成熟するとともに，分野が細かく細分化され，また，処理対象となるものも，新聞以外に論文，特許，WWW 上のテキストなど多岐にわたり，さらに，応用システムもさまざまなものが生まれつつある。そして，自然言語処理は，現在では，WWW の普及とともに，ネットワーク社会の基盤を支える重要な情報技術の一つとなっているといえる。

これまでの自然言語処理に関する専門書は，自然言語処理全般を広く浅く扱う教科書（入門書）以外には，情報検索，テキスト要約などを扱う，わずかの書籍が出版されているだけという状況であった。この現状を鑑みるに，読者は，「実際にいま役に立つ本」,「いまの話題に即した本」を求めているのではないかと推測される。そこで，これまでの自然言語処理に関する専門書では扱われておらず，なおかつ,「いま重要と考えられ，今後もその重要さが変わらない」と考えられるテーマを扱った書籍の出版を企画することになった。

このような背景の下生まれた「自然言語処理」シリーズの構成を以下に示す。

1. 自然言語処理で利用される，統計的手法，機械学習手法などを広く扱う
   近年の自然言語処理は，コーパスに基づき，統計的手法あるいは機械学習手法を用いて，規則なり知識を自動獲得し，それを用いた処理を行うという手法を採用することが一般的になってきている。現状多くの研究者は，他の先端的な研究者の論文などを参考に，それらの統計的手法，機械学習手法に関する知識を得ており，体系的な知識を得る手がかりに欠けている。そこで，そのような統計的，機械学習手法に関する体系的知識を与える専門書が必要と感じている。
2. 情報検索，テキスト要約などと並ぶ，自然言語処理の応用を扱う
   自然言語処理分野も歴史を重ね，技術もある程度成熟し，実際に使えるシステム，技術として世の中に少しずつ流通するようになってきている

ものも出てきている。そのようなシステム，技術として，検索エンジン，要約システムなどがあり，それらに関する書籍も出版されるようになってきている。これらと同様に，近年実用化され，また，注目を集めている技術として，情報抽出，対話システムなどがあり，これらの技術に関する書籍の必要性を感じている。

3. 処理対象が新しい自然言語処理を扱う

自然言語処理の対象とするテキストは，近年多様化し始めており，その中でも，注目を集めているコンテンツに，特許（知的財産），WWW 上のテキストが挙げられる。これらを対象とした自然言語処理は，その処理結果により有用な情報が得られる可能性が高いことから，研究者が加速度的に増加し始めている。しかし，これらのテキストを対象とした自然言語処理は，これまでの自然言語処理と異なる点が多く，これまでの書籍で扱われていない内容が多い。

4. 自然言語処理の要素技術を扱う

形態素解析，構文解析，意味解析，談話解析など，自然言語処理の要素技術については，教科書の中で取り上げられることは多いが，技術が成熟しつつあるにもかかわらず，これまで技術の現状を詳細に説明する専門書が書かれることは少なかった。これらの技術を学びたいと思う研究者は，実際の論文を頼らざるを得なかったというのが現状ではないかと考える。

本シリーズの構成を述べてきたが，この構成は現在の仮のものであることを最後に付記しておきたい。今後これらの候補も含め，新たな書籍が本シリーズに加わり，本シリーズがさらに充実したものとなることを祈っている。

本シリーズは，その分野の第一人者の方々に各書籍の執筆をご快諾願えたことで，成功への最初の一歩を踏み出せたのではないかと思っている。シリーズの書籍が，読者がその分野での研究を始める上で役に立ち，また，実際のシステム開発の上で参考になるとしたら，この企画を始めたものとして望外の幸せである。最後に，このような画期的な企画にご賛同下さり，実現に向けた労をとって下さったコロナ社の各氏に感謝したい。

2013 年 12 月

監修者　奥 村　　学

ま　え　が　き

医療分野の情報処理は加速している。電子カルテの実用化により医学・診療情報の蓄積が急速に進みつつある。加えて，それを処理する機械学習やディープラーニングなど新しいデータ処理技術も研究が進んでいる。間違いなく，医療分野の情報処理の高度化は，今後の社会変革のトリガーになるであろう。そして，その基盤を支える技術として，自然言語処理は必須の要素となるはずである。

しかし，現在まで，医療分野の言語処理に関する書籍は刊行されていない。医療分野の言語処理を進めていくうえで必要となる情報は 2 種類ある。まず，情報処理研究者にとっては，医療分野にどのような研究材料があり，どのようなリソースが利用可能で，なにを目指せばよいかといったことが重要であろう。一方，医療情報処理に従事してきた研究者や実務者にとっては，どのような自然言語処理技術がどの程度の水準にあるのか？　ある目的のためになにを使えばよいのか？　といったことが必要となるであろう。本書は，前者を想定している。まず，医療においてこれまで開発されたリソースを紹介している。リソースは，電子カルテ文章だけでなく，ウェブテキストや患者の語りなど，医療文章以外のテキストも対象としている。なかなかなじみのない医療データを紹介するために，実際のデータに近いデータや，書式も含めるように工夫した。手法については，研究事例をベースとしたが，単なる研究紹介にとどまらず，教科書としての使用も可能なように，基本を押さえた内容となるように配慮している。本書が医療分野の言語処理の発展の一助となることを祈念している。

なお，本書の執筆にあたっては多くの人々の助力を得た。特に，奈良先端科学技術大学院大学情報科学研究科の若宮翔子博士研究員，伊藤　薫研究員，金子雅美技術補佐員，山本英弥君には丁寧なチェックをしていただいたことを深

く感謝する。また，慶應義塾大学大学院薬学研究科臼井美紗さんには医薬品情報学に関する指導，東京大学医学部附属病院河添悦昌講師，篠原恵美子特任助教には医療情報学に関する指導をいただき，感謝する。

2017 年 5 月

荒牧 英治

目　　次

1. 医療情報の利活用とは

1.1 医療情報学の歴史 …… 2
1.2 病院内テキストの利活用 …… 4
1.3 パブリックデータの利活用 …… 5
1.4 プライベートデータの利活用 …… 6
1.5 各 国 の 動 向 …… 7
1.6 今 後 の 動 向 …… 8

2. 利用可能なリソース・ツール

2.1 辞書・シソーラス・オントロジー …… 10
2.1.1 人間のための辞書 …… 10
2.1.2 オントロジーとシソーラス …… 12
2.1.3 バイオインフォマティクス・オントロジーと
クリニカル・オントロジー …… 13
2.1.4 SNOMED-CT …… 14
2.1.5 ICD-10 …… 17
2.1.6 MeSH …… 19
2.1.7 MedDRA …… 20
2.1.8 UMLS …… 21
2.1.9 医薬品に関するリソース …… 23

2.1.10 検査に関するリソース …… 28
2.1.11 治療・処置に関するリソース …… 30
2.1.12 標準病名マスター …… 31
2.2 その他の辞書・リソース …… 32
2.3 コ ー パ ス …… 35
2.3.1 i2b2 NLP コーパス …… 35
2.3.2 GSK 診療録コーパス …… 36
2.3.3 NTCIR MedNLP コーパス …… 39
2.4 言 語 ツ ー ル …… 40

3. 病院内テキスト

3.1 病院内テキストとは …… 42
3.1.1 診 療 録 …… 44
3.1.2 サ マ リ …… 48
3.1.3 看 護 記 録 …… 48
3.1.4 読影レポート・病理レポート …… 50
3.1.5 手術記録・麻酔記録 …… 52
3.1.6 説明書・同意書 …… 52
3.1.7 その他のコメディカル文書 …… 52
3.1.8 レセプトデータ …… 52
3.1.9 有害事象報告 …… 56
3.1.10 副 作 用 報 告 …… 57
3.1.11 救命救急文書 …… 59
3.2 おもな研究課題 …… 60
3.2.1 固有表現認識ベースの匿名化 …… 61
3.2.2 プライバシー保護マイニング・ベースの匿名化 …… 66

3.2.3 自動コーディング …… 67
3.2.4 患者情報抽出 …… 69
3.2.5 診断支援・自動診断 …… 72
3.2.6 標準化（表記ゆれ吸収） …… 75
3.2.7 副作用シグナルの自動検出 …… 77
3.2.8 入力支援 …… 81
3.2.9 NTCIR MedNLP シリーズ …… 82
3.3 カルテテキストへのアノテーション …… 83
3.4 アノテーションにおける諸問題 …… 90
3.5 倫理申請 …… 95

4. パブリックデータ：公開テキストの医療言語処理

4.1 さまざまな公開テキスト …… 97
4.1.1 学術論文 …… 97
4.1.2 研究スタイル …… 99
4.1.3 臨床試験登録情報 …… 103
4.1.4 NDB …… 105
4.1.5 コホートデータ …… 105
4.1.6 ソーシャルメディアのデータ …… 106
4.1.7 バイオ NLP コーパス …… 109
4.1.8 その他のデータ …… 110
4.2 おもな研究課題 …… 111
4.2.1 論文検索 …… 111
4.2.2 タンパク質相互作用抽出 …… 114
4.2.3 構造を考慮した検索高速化 …… 115
4.2.4 感染症サーベイランス …… 116

4.3 実アプリケーション …… 121
4.4 ソーシャルメディア・データのラベル付け …… 124

5. プライベートデータ：患者テキストの医療言語処理

5.1 患者の記述するテキスト …… 130
5.1.1 疾患別の患者テキスト …… 131
5.1.2 情報収集源としての患者テキスト …… 132
5.1.3 三つのアプローチ …… 135
5.2 QOL アプローチ …… 137
5.3 教育アプローチ …… 138
5.4 研究アプローチ …… 140
5.5 おもな研究課題 …… 141
5.5.1 表記ゆれ吸収 …… 141
5.5.2 ウェブ情報の信頼性 …… 142
5.5.3 高齢者の孤立を防ぐコミュニケーションツール …… 148

6. これからの医療言語処理研究

6.1 研究を始めるにあたって …… 150
6.1.1 ジャーナル …… 150
6.1.2 国際会議 …… 151
6.2 今後の展望 …… 151

引用・参考文献 …… 153
索 引 …… 165

1 医療情報の利活用とは

医療現場で生成される多様なデータの相当な部分は自然言語文であり，今後もそれはただちに変わりそうにない。医療データの利活用には，情報発生源である診療へ応用する利用（一次利用），および，その一次利用の結果，集積されたデータを研究や政府の施策に活かすといった別目的での利用（二次利用）の二つがある。現在，盛んに利活用が叫ばれているのは，後者の二次利用である。これまで医療データの二次利用は，健診データや診断群分類別包括支払い制（diagnosis procedure combination/per-diem payment system, DPC/PDPS）の診療報酬データなど，比較的構造化されたデータがおもな材料であった。

しかし，最近では，ビッグデータ解析の流れを受けて，より大規模，かつ非構造化されたデータを扱う方向へ発展しつつある。その大規模，かつ非構造的なデータの代表が自然言語文である。本書は，この医療分野における自然言語データの利活用を扱うものである。その利活用の方向はつぎの三つに集約される。第一に，大きな動向は，診療録（電子カルテ）に代表される医師が日常診療で残すデータの利活用を目指す方向である（病院内テキストの利活用）。もう一つの大きな動向は，論文やウェブ情報などの公開されているデータを残す方法である（パブリックデータの利活用）。最後に，この数年ほどの間に急速に注目を集めているのが，患者がソーシャルメディアや患者会などを通じてやり取りするきわめてプライベートなデータを扱う方向である（プライベートデータの利活用）。本書は，まず，これまでの医療情報学の歴史を振り返ることからはじめ，それぞれのアプローチが生まれるに至った背景を解説する。

1.1 医療情報学の歴史

医療情報学の歴史は 1960 年代に端を発する。まずは，会計などの効率化からコンピュータが利用され始め，診療行為や薬剤などの情報をシステムに入力することにより，医療費の計算が自動化された。この流れは，会計にとどまらず，自動分析装置を統合的に制御する検査システムや，オーダリングシステム（オーダエントリーシステムともいう）といわれる検査や処方の依頼（オーダ）を行うシステムなど，次々と業務システムが電子化されていく。しかし，これらはバックエンドの業務システムであり，一般の人々の目に触れることは滅多にない。人々が病院の電子化と聞いてまっさきに連想するのは電子カルテ，正確には**診療録**（health record または medical record：医師が患者と対面して記述する記録）であるが，この電子カルテの普及には時間を要した。これは，初期導入時にかかるコストの問題，技術者の不足といった通常の IT システムの問題だけでなく，コンピュータシステムへの記録が，医師法による法的根拠のある文書かどうか，など，法改正も含めた問題となったからである。しかし，1999 年，一定の基準を満たした電子媒体への保存であれば診療録として認められるという法改正が行われた。これにより，日本において，初めて電子カルテが利用可能となり，以降，急速に電子カルテが普及しつつある。現在では，オーダリングシステムは一般病院のほとんどに普及し（**表 1.1**），電子カルテシステムも，大規模病院のほとんどと一部の中小規模病院に普及し始めている（**表 1.2**）。

表 1.1　オーダリングシステム普及の推移〔厚生労働省医療施設調査より〕

年	一般病院	臨床規模別		
		400 床以上	200～399 床	200 床未満
2008 年 （平成 20 年）	31.7％ (2 448/7 714)	82.4％ (593/720)	54.0％ (745/1 380)	19.8％ (1 110/5 610)
2011 年 （平成 23 年）	39.3％ (2 913/7 410)	86.6％ (606/700)	62.8％ (827/1 317)	27.4％ (1 480/5 393)
2014 年 （平成 26 年）	47.7％ (3 539/7 426)	89.7％ (637/710)	70.6％ (946/1 340)	36.4％ (1 956/5 376)

表 1.2 電子カルテシステム普及の推移〔厚生労働省医療施設調査より〕

年	一般病院	臨床規模別			一般診療所
		400 床以上	200～399 床	200 床未満	
2008 年 （平成 20 年）	14.2％ (1 092/7 714)	38.8％ (279/720)	22.7％ (313/1 380)	8.9％ (500/5 614)	14.7％ (14 602/99 083)
2011 年 （平成 23 年）	21.9％ (1 620/7 410)	57.3％ (401/700)	33.4％ (440/1 317)	14.4％ (779/5 393)	21.2％ (20 797/98 004)
2014 年 （平成 26 年）	34.2％ (2 542/7 426)	77.5％ (550/710)	50.9％ (682/1 340)	24.4％ (1 310/5 376)	35.0％ (35 178/100 461)

この電子カルテの普及によってなにが変わるのであろうか？ 電子カルテ導入の第一の目的は，病院の運営コストの削減である。大病院，例えば，大学病院規模であれば，A4 サイズの用紙のような定型用紙のものだけでなく，レントゲン写真，伝票といったさまざまな体裁，様式をとる書類を大量に扱う。これが電子化されるだけでも大幅なコスト削減が期待される。さらには，これらの保管スペースが不要になる。つぎに，一部のデータ入力作業，集計作業が短縮されることで医療事務の人件費が削減される。こういった恩恵は，病院側が受けるものだけでなく，患者側にとっても，業務の効率化などにより，患者の待ち時間が短縮されるなどの効果がある。さらに，電子化されたことにより，ミスや誤読が減少し，安全性の向上も期待される。

しかし，本当の恩恵は上記にとどまらない。電子化された情報が蓄積されることにより，これまで不可能であった，大規模，かつ網羅的な患者動態の追跡が可能となり，新たな医学研究の材料となりつつある。例えば，医薬品，医療機器副作用などの安全に関わる情報の収集，疾患に関する疫学的情報，これらのような膨大な労力をかけて行われてきた調査研究がより大規模，かつ容易に実行できる。さらに，診断支援，類似症例検索といったこれまで不可能であった医療情報サービスも構築可能となる。

このような背景を鑑みれば，電子カルテというインフラが整ってはじめて，医療情報学は，単なるシステム開発を超えて，より積極的に医療に関わる基盤を得たといえる。

さらに，近年は，この動きを加速させる流れも多方面から起こりつつある。

一つは，研究論文のアーカイブや，患者がソーシャルメディアなどに主体的に残すテキストといったパブリックデータであり，オープンジャーナル化などアカデミアのオープン化の流れを受けて新たな医療研究の材料になりつつある。

もう一つは，患者が患者会などを通じてやり取りする電子的なデータであり，ここではプライベートデータと呼ぶ。ソーシャルメディアやスマートフォンの普及とともに，患者が主体となってデータをアーカイブしつつある。それでは，これら三つの材料，病院内データ（以降，病院内テキスト），パブリックデータ，プライベートデータについて，それぞれの動向を簡単にまとめる。

1.2　病院内テキストの利活用

病院内テキストの利活用では，電子カルテ（退院サマリや診療録などの診療記録テキスト）を扱う研究が中心となる。病院，特に，大規模病院で日常的にやり取りされる書類の数は膨大な数になり，大学病院規模では 1 か月に 20 万以上もやり取りされるとの統計がある。これらは基本的には，病院運営の業務のために使用されるが，蓄積されたカルテテキストを利用することでかつてない大規模な研究が実施可能となる。中でも，診療録（以降，電子カルテ）の記述は，入院時初期記録から退院サマリまでの患者の全体像が生々しく，かつあまねく表現されており，最も利活用が期待されているデータの一つである。

このデータの潜在的な活用は二つに大別される。① まずは，大量の医療情報を分析・利活用し，新たな臨床研究を推進することが挙げられる。② もう一つは，健康医療政策に資する統計データの収集に貢献することも期待される。

ただし，電子カルテの記述内容の具体については，医療従事者に一任されてきた面が多く，その結果として，非文法的，かつ断片化した表現が多く含まれている。このため，病院ごとに異なった形で行われているカルテから，医療情報（例えば，傷病名や愁訴，検査，治療に関する記述）を抽出し，標準化することは難しく，新たな自然言語処理（natural language processing，NLP）の課題として活発に研究されている。

同時に，すでに大量に蓄積されている既存の電子カルテ情報に対して処理する（後ろ向き解析）だけでなく，新しく電子カルテに医師などが入力する際に，裏側で標準化された結果を提示する処理（前向き解析）に関する研究も行われている。

最後に，これらの研究成果は，究極的には，異なる電子カルテベンダーによって開発されたさまざまな種類やバージョンのカルテに導入される必要があり，どのような形で実際のカルテシステムに組み込むかという課題がある。

1.3　パブリックデータの利活用

パブリックデータ（公開された情報）の利活用としては，おもに医学論文やウェブテキストが材料とされている。医学論文の蓄積は膨大なものになりつつあり，例えば，代表的な医学論文データベースである PubMed には現在すでに 2 600 万件[†1]を超える論文のアブストラクトが登録されている。中でも，特にがんおよびゲノム関係の論文の数は著しく，がんに関する論文だけでも 2014 年だけで 20 万報を超えている。この膨大な成果を一人の専門家，または，少数の研究グループが把握することは困難となりつつあり，学問領域を検索する技術，および，俯瞰（ふかん）する技術が研究されている。この検索においては，Google[†2]のような汎用の検索と異なり，ドメインに特化した検索要求に応えられることが望ましい。例えば，タンパク質の作用の関係，医薬品と副作用の関係，疾患と症状の関係など，個別のニーズに特化した検索技術が研究されている。このような高度な検索を実現するためには，用語の標準化だけでなく，用語どうしの関係を捉えるために，関係抽出や構文解析といった深い言語解析が必要となる。さらに，解析の基盤となる大規模な**コーパス**（corpus：用語の関係が付与されたテキストデータ）も必須である。英語圏においては，この研究環境が，GENIA

†1　2017 年 2 月現在。

†2　本書で使用している会社名，製品名は，一般に各社の商標または登録商標です。本書では ® と ™ は明記していません。

索　　引

【あ】

アーカイブ 4
アノテーション 38
アノテーション付きコーパス 116
アブストラクト 5
アメリカ化学会 28
アメリカ国立医学図書館 19
アメリカ臨床病理医協会 14
誤り訂正タスク 76
アルゴリズム 16
アレルギーの有無 44
安全性定期報告 20

【い】

言い換え 67
医科入院レセプト 105
医師国家試験 73
医師国家試験問題 39
医師法 2
一次利用 1
一般化 66
一般化操作 67
一般化単位 67
一般事実表現 126
意味関係獲得 76
意味構造検索 112
医薬品 83
医薬品規制調和国際会議 20
医薬品添付文書 23
医薬品販売量 110
医薬品副作用・感染症報告 20
医薬品 HOT コードマスター 31
医療機器副作用 3
医療健康情報認証機構 145
医療言語処理タスク 39
医療コーパス 35
医療従事者 4
入れ子 116
飲酒歴 69
インパクトファクター 150
インフラ 3
インフルくん 122

【う】

ウェブテキスト 5
後ろ向き解析 5
後ろ向きコホート研究 102

【え】

疫学的情報 3
疫学モデル 119
エビデンスレベル 101

【お】

オーダエントリーシステム 2
オーダリングシステム 2
オントロジー 7

【か】

外国語表記 75
ガイドラインアプローチ 143
概念 12
概念志向用語集 14
回復表現 127
回復率 120
外来 42
外来患者対応 42
書き間違い 76
書き間違い訂正タスク 76
カゼミル・プラス 122
画像検査マスター 31
下層語 20
家族歴 44
カルテ 44
含意表現 127
関係抽出 5, 78
看護記録 43, 48
看護実践用語標準マスター 31
患者会 1
患者喫煙状態 69
患者情報抽出 69
患者説明資料 95
患者テキスト 130
患者動態 3
患者報告アウトカム 130
患者 SNS 6, 7, 131
感染症サーベイランス 117
感染症定期報告 20
感染率 120
関連 12

【き】

既往歴 44
機械学習 8
器官別大分類 20
既喫煙者 69, 71
基礎系論文 98
喫煙・飲酒 44
喫煙者 70
喫煙状態 69

喫煙歴 69
基本語 20
客観的情報 47
救急救命士活動記録 59
救命救急外来診療録 59
救命救急文書 59
教育アプローチ 136, 138
教育用コンテンツ 138
教育用電子カルテ共同利用協議会 36
狭義の症状 80

【く】

クリニカル・オントロジー 13
グレード 101

【け】

形態素解析 33
ケースコントロール研究 103
ケースシリーズ研究 103
決定木 68
現喫煙者 69
研究アプローチ 136, 140
検査 43
検索クエリ 111
検索エンジン 111
検査システム 2
健診データ 1, 9
原著論文 97
現病歴 44

【こ】

高位グループ語 20
高位語 20
広義の症状 80
厚生省コード 24
構文解析 5, 71
コーディング 18, 67
コーパス 5
コールセンターログ 110
語義曖昧性解消 76
国立医学図書館 21
国立情報学研究所 39
個別医薬品コード 25
コホート研究 102
コメディカル 52
固有表現認識 62

【さ】

財団法人日本医薬情報センター 105
サマリ 43, 48

【し】

シェアードタスク 8
歯科手術・処置マスター 31
歯科病名マスター 31
歯科レセプト 105
識別タスク 76
軸 18
時空間 124
シグナル 77
指示 43
事実性 124
事実性の有無 84
事実モダリティ的表現 125
静岡分類 133
システマティックレビュー 99, 102
時制 128
シソーラス 10, 13
悉皆データ 105
疾患名 37
自動診断 72
主観的情報 47
手術記録 52
手術・処置 43
手術・処置マスター 31
主訴 44
出版バイアス 104
準ランダム化比較試験 102
上位下位関係獲得 76
上位下位関係 13
紹介状 43
症状 83
症状所見マスター 31
症状名 37
承認番号 25
情報検索尺度 72
情報抽出 71
症例くん 113
症例検索 113
症例報告 99, 103
初診 43
人工知能 72
診断群分類別包括支払い制 1, 52
診断支援 3, 72
人名 38
信頼性 145
診療報酬コード 18
診療報酬データ 7, 9
診療報酬点数表 30
診療録 1, 2, 44

【せ】

生活の質 48
生成タスク 76
性別 38
世界保健機関 17
説明書 52
専門医 99
専門家の意見 103

【そ】

ソーシャルセンサ 106
ソーシャルメディア 1, 106
粗化 67

【た】

ターミノロジー 8
退院サマリ 4, 48
退院時要約 48
タイポ 39
タンパク質相互作用抽出 114

【ち】

治験薬副作用・感染症症例報告 20
超過死亡概念 116
調剤レセプト 105
直接表現 124
治療方針 47

【つ】

ツイートフル 122
通常の流行時 118
ツリー構造 13

【て】

テキストマイニング 136
出来高払い方式 52
電子カルテ 1
電子カルテテキスト 35
電子カルテベンダー 5
テンプレート化 61

【と】

同意書 43, 52
同格 87
同義語獲得タスク 76
統合（一体化）医学用語システム 21
当事者 117
当事者ブログ 131
闘病記 131
闘病記録 6
闘病生活 6
登録データベース 95
読影レポート 50
匿名化 61, 62
独立行政法人医薬品医療機器総合機構 58
突発的な流行時 118
ドメイン 5

【な】

ナイーブベイズ分類器 68
ナラティブデータ 6

【に】

二次利用 1
日時 38
日英医学翻訳辞書 16
日本医師会 95
日本医師会治験促進センター 105
日本医薬情報センター 28, 95, 103
日本インターネット医療協議会 144
日本標準商品分類番号 23
日本臨床検査医学会 28
入院外レセプト 105
入院患者対応 42
入退院関係 43
認証アプローチ 143
認定医 99

【ね】

ネガティブ・イベント 84
年齢 36

【は】

パーソナライゼーション 61
バイオインフォマティクス・オントロジー 13
バイオ自然言語処理 109
バイオ NLP 109
場所名 37
バックエンド 2
発現 34
パブリックデータ 1, 5

【ひ】

比較臨床試験 102
非喫煙者 70, 71
非事実モダリティ的表現 125
ビッグデータ化 9
ビッグデータ解析 1
比喩表現・誤字 126
病院名 37
評価診断 47
表記 12
表記のゆれ 75
表記ゆれ吸収 75
標準医薬品コード 31
標準化 4, 75
標準病名 31
標準病名マスター 31, 68
病棟 42
病理レポート 50
非ランダム化比較試験 102

【ふ】

フォーカス・チャーティング形式 47
深い言語解析 5
不均衡データ 81
副作用 57, 77
副作用報告 57
不明 70
プライベートデータ 1
文書分類タスク 71
分類タスク 76

【ほ】

北米放射線学会 33
保険請求 52
ポジティブ・イベント 84
翻字 76
翻字タスク 76

【ま】

前向き解析 5
前向きコホート研究 102
麻酔記録 52
マッピング 17
マルチラベリングタスク 68

【み】

みんくす 139
みんなの花粉症なう 121

【め】
メタシソーラス 21
【も】
模擬病歴報告 39
モダリティ 84
【や】
薬物有害反応 57
薬価基準収載医薬品コード 24, 26
薬価コード 24
薬効分類番号 12, 24
薬効レベル 12
病の語り 6, 131
【ゆ】
有害事象 20, 56
有害事象報告書 57
有害反応 56
ユーザ生成コンテンツ 6
ユーザ・バイアス 107
【よ】
用例 68
用例ベース 68
【ら】
ライフサイエンス辞書 33
ライフタイムマネジメントセンター 139
ライフパレット 140
ライブラリ化 6
ランダム化比較試験 102
【り】
略語展開 76
略語展開タスク 76
粒度 13
領域代数 116
臨床医学オントロジー 32
臨床系論文 98
臨床研究 52
臨床検査マスター 31
臨床試験登録・公開システム 105
臨床試験登録システム 103
臨床試験登録情報 103
倫理規約 145
倫理的問題 74
【る】
類似症例検索 3
【れ】
レセプト情報・特定健診等情報データベース 105
レセプトデータ 52
レセプト電算処理システム用コード 26
レビュー論文 99
連携 43
【ろ】
論文 97
【わ】
ワシントン大学 33
【A】
ACE 78
ACS 28
ADR 57
adverse drug reaction 57
adverse event 56
adverse reaction 56
AE 56
age 36
American Chemical Society 28
American Medical Informatics Association 151
AMIA 151
Apache cTAKES 40
apposition 87
AR 56
articles 97
assessment 47
ATC コード 12, 27
【B】
BioCreative 78
BMJ 98
BM25 72
British Medical Journal 98
【C】
CAS 登録番号 28
case control study 103
case report 99
CASP 147
CLEF eHealth 8
CLEF e-Health 35
ClinicalTrials.gov 104
Clinical Trials Listings Service 104
cohort study 102
ComeJisyo 33
complaint 37
complete enumeration data 105
concept 12
conditional random field 65
corpus 5
CRF 65

current smoker 69

【D】

DAR 形式 47
description 12
de-identification 62
diagnosis procedure combination 52
diagnosis procedure combination/per-diem payment system 1
DIPEx 7, 139
DIPEx Japan 7
DIPEx-Japan 139
DISCERN 146
DNA 情報 34
document classification 71
DPC 方式 52
DPC レセプト 105
DPC/PDPS 1

【E】

epidemic 時 118
EU 35
EudraCT 105
evaluation 47
expert determination 61

【F】

FC 形式 47
FMA 33
Foundational Model of Anatomy Ontology 33

【G】

Gene Ontology 34
GENIA プロジェクト 6
genotype 34
Google 検索クエリ 121
Google Flu Trends 121
GSK 診療録コーパス 36
GS1 25

【H】

HealthLink 111
Health Insurance Portability and Accountability Act 7, 61
health record 2
high level group term 20
high level term 20
HIPAA 7, 61
HLGT 20
HLT 20
HONcode 145
hospital 37
HOT コード 12, 26
HPO 34
Human Phenotype Ontology 34

【I】

ICD コーディング 18, 67
ICD コード 67
ICD-10 12
ICD-10 コード 12
ICD-10 対応標準病名マスター 31
ICD-9-AM 17
ICD-9-CM 17
ICNP 33
IER 形式 47
IF 150
IHTSDO 14
imbalanced data 81
IMIA 151
IMRAD 98
incognito 法 67
information retrieval 71
International Classification for Nursing Practice 33
International Health Terminology Standards Development Organization 14
intervention 47
Introduction, Methods, Results and Discussion 98
IR 71
IT 化 8
i2b2 NLP コーパス 35
i2b2 NLP Challenge 35, 151
i2b2 NLP Smoking Challenge 69

【J】

JACHI 145
JAMA 98
JAN コード 12, 25, 26
JAPIC 28, 105
JAPIC コード 28
JapicCTI 95
JAPIC CTI 103
JIMA 144
JIS 化 25
JLAC10 28
JMACCT 95
JMACCT CTR 103
Journal of the American Medical Association 98
J-RARE.net 140

【K】

k-匿名化 66

【L】

LINE 106
Linguistic String Project 40
LLT 20
location 37
Logical Observation Identifiers Names and Codes 29
Logoscope 72
LOINC 29
lowest level term 20
LSP 40

【M】

Mayo clinical Text Analysis and Knowledge Extraction System *41*
MedDRA *20, 58*
medical literature analysis and retrieval system online *97*
medical record *2*
Medical Subject Headings *19*
MEDIE *112*
MEDINFO *151*
MEDIS-DC *31*
MEDLARS Online *97*
MedLEE *40*
MEDLINE *19, 97, 109*
MEDNLP *82*
MedNLP *39*
MedNLP-DOC *69*
MeSH *19*
MeSH ターム *19, 111*
Mondrian 法 *67*
MUC *78*
MYCIN *72*

【N】

named entity recognition *62*
national data base *105*
National Library of Medicine *21*
NCBO BioPortal *13*
NDB *105*
negative *39*
NEJM *98*
NER *62*
NHS Trusts Clinical Trials Register *104*
NII *39*
non-randomized controlled trial *102*
non-smoker *70*
NTCIR *39, 82*
NTCIR MedNLP *7*
NTCIR12 *69*

【O】

objective *47*

【P】

pandemic 時 *118*
papers *97*
past smoker *69*
PatientsLikeMe *7, 140*
patient narrative *131*
patient reported outcome *140*
patient reported outcomes *130*
PATO *34*
person *38*
Pharmaceuticals and Medical Devices Agency *58*
phenotype *34*
Phenotypic Quality Ontology *34*
PhRMA 臨床試験結果データベース *105*
PINACO *113*
plan *47*
PMDA *58*
positive *39*
preferred term *20*
PRO *140*
prospective cohort study *102*
PT *20*
PubChem CID *28*
publication bias *104*
PubMed *5, 97, 109*

【Q】

QOL *7, 48*
QOL アプローチ *135, 137*
quality of life *7*
QUICK *147*

【R】

RadLEX *33*
randomized controlled trial *102*
RCT *102*
region algebra *116*
relationship *12*
relation extraction *78*
retrospective cohort study *102*
revision *47*

【S】

safe harbor *61*
SE *57*
SemEval *78*
sex *38*
side effect *57*
SIR モデル *120*
smoker *70*
SNOMED-CT *14*
SNS *7*
SOAP 形式 *47*
SOC *20*
social media *106*
social sensor *106*
subjective *47*
support vector machines *65*
suspicion *39*
SVM *65*
systematic review *102*
Systematized Nomenclature of Medicine-Clinical Terms *14*
system organ class *20*

【T】

TEXT2TABLE *80*

The Health on the Net Foundation Code of Conduct（HONcode）for medical and health Web sites *145*
The International Standard Randomized Controlled Trial Number Register *104*
The ISRCTN resistry *104*
The New England Journal of Medicine *98*
time *38*
TOBYO *140*
top-down 法 *67*
transliteration *76*
Twitter *6, 106*

【U】

UGC *6*
UMIN *105*
UMIN 臨床試験登録システム *95, 103, 104*
UMIN Clinical Trials Registry *103*
UMIN-CTR *103, 104*
UMLS *12, 21, 41*
Unified Medical Language System *21*
unknown *70*
user generated contents *6*

【W】

WHO *17*
word sense disambiguation *76*
World Congress on Medical and Health Informatics *151*
WSD *76*

【Y】

YJ コード *24, 26*

監修者・著者略歴

奥村　学（おくむら　まなぶ）

1984年　東京工業大学工学部情報工学科卒業
1989年　東京工業大学大学院博士課程
　　修了（情報工学専攻）
　　工学博士
1989年　東京工業大学助手
1992年　北陸先端科学技術大学院大学助教授
2000年　東京工業大学助教授
2007年　東京工業大学准教授
2009年　東京工業大学教授
　　現在に至る

荒牧　英治（あらまき　えいじ）

2000年　京都大学総合人間学部基礎科学科卒業
2002年　京都大学大学院情報学研究科修士課程
　　修了（知能情報学専攻）
2005年　東京大学大学院情報理工系研究科博士
　　課程修了（電子情報学専攻）
　　博士（情報理工学）
2005年　東京大学医学部附属病院特任助教
2008年　東京大学知の構造化センター特任講師
2011年　京都大学デザイン学ユニット特定准教授
2015年　奈良先端科学技術大学院大学
　　特任准教授
　　現在に至る

医療言語処理

Introduction of Medical Natural Language Proccessing

2017年8月25日　初版第1刷発行

検印省略

監修者　奥村　学
著　者　荒牧　英治
発行者　株式会社　コロナ社
　　代表者　牛来真也
印刷所　三美印刷株式会社
製本所　有限会社　愛千製本所

112–0011　東京都文京区千石4–46–10
発行所　株式会社　**コロナ社**
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話(03)3941–3131(代)
ホームページ　http://www.coronasha.co.jp

ISBN 978–4–339–02762–4　C3355　Printed in Japan　（森岡）