
一般口演

一般口演5 知識工学

2017年11月21日(火) 08:30 ～ 10:00 G会場 (10F 会議室1006-1007)

[2-G-1-OP5-1] 医療文書分析の目的限局性に基づく単語階層の構成法の検討

松尾 亮輔, Ho Tu Bao, 池田 満, 田中 孝治, 陳 巍（北陸先端科学技術大学院大学）

電子カルテの二次利用の整備が進み、カルテ内の膨大な医療情報の利活用への期待が高まっている。診療支援のために、医療文書を用いてデータマイニングによる予測や検索などの分析がされている。文書进行分析するには、文書を数値で計算可能な形式に表現する。医学文書进行分析の際は、医学オントロジーを活用して、文書内の単語の概念に着目することで、類義語や同義語の同定や単語タイプの識別により、目的志向の重み付け手法が開発され、検索や分類などに用いられている。

先行研究において、医学単語集合を分析目的独立のものから依存性の高いものへと順に配置した単語階層を用いて、分析目的限局性の重み付けをすることが、分析精度の向上に有効であることを実験的に示した。特に階層の上位から順に、重み付けに用いる単語の種類を増やすと、分析目的独立の単語集合と目的依存の単語集合の境界で、分析精度の向上が認められた。よって、目的独立単語の下位に目的依存単語を配置する汎用手法を構成できることが示唆された。

本研究で構成した手法は以下である。入力は医療文書内の単語集合である。分析目的と独立する単語集合は、最上位で非医学単語集合と医学単語集合からなり、医学単語集合は ICD-10医学単語集合と非 ICD-10医学単語集合からなる。出力は分析に用いられる単語の階層構造である。提案手法の処理手順は、ICD-10単語が持つ複数の概念について、分析目的との関連により、様々な分析目的に依存する単語集合を構成する。さらに、単語の共起により学習される類似性に基づき、非 ICD-10医学単語集合と非医学単語集合から、分析目的による単語の集まりに類似する単語集合を構成する。

医学知識基盤が充実し、機械学習技術が進展する中で、目的志向の重み付けの観点から、既存知識の組み合わせと、機械学習の活用により、新たに分析目的による階層的な単語分類体系を作ることによって、医療ビッグデータ分析による知識獲得支援が期待できる。

医療文書分析の目的限局性に基づく単語階層の構成法の検討

松尾 亮輔^{*1}、Ho Tu Bao^{*1}、池田 満^{*1}、田中 孝治^{*1}、陳 巍^{*1}

^{*1} 北陸先端科学技術大学院大学

Investigation of a construction method of term hierarchy based on the analytical purpose of medical document analysis

Ryosuke Matsuo^{*1}, Tu Bao Ho^{*1}, Mitsuru Ikeda^{*1}, Koji Tanaka^{*1}, Wei Chen^{*1}

^{*1} Japan Advanced Institute of Science and Technology

The secondary use of electronic medical records (EMRs) has been pursued in Japan and it is expected to utilize a huge amount of medical information in EMRs. Term weighting is essentially significant for medical document analysis which especially requires natural language processing techniques because of the nature of medical documents. Although there are many term weighting methods depending on analytical purposes, with the methodology supporting the pre-processing, enhancing knowledge acquisition support by big data analysis in medical domain is accomplished. To this end, the objective of this paper is to investigate a construction method of term hierarchy which composes of terms' set based on the analytical purpose of medical document analysis. We considered term hierarchy which composes of independent and dependent terms' set for analytical purposes, by exploiting medical knowledge UMLS, ICD-10, a ranking corresponding to ICD-10 and a machine learning technique. We especially employed Word2Vec to consider similar dependent terms' set based on the proximity relationship, to deal with the problem of ICD-10 where the codes are assigned to diseases, despite not only terms regarding diseases but also terms regarding drugs, tests and procedures are contained in medical documents. We also considered the exploitation of the term hierarchy for supporting analytical purpose-based term weighting.

Keywords: Natural language processing, Knowledge engineering, Artificial intelligence, Knowledge representation, ICD.

1. 結論

現在、日本は電子カルテを二次利用するための整備が進んでおり、カルテ内の膨大な医療情報の利活用への期待が高まっている。これまで、診療支援のために、医療文書を用いてデータマイニングによる予測や検索などの分析がされている。文書を分析するには、通常、単語の重み付けにより、文書を数値で計算可能な形式に表現する。医学文書を分析する際は、Medical Subject Headings (MeSH)¹⁾や Unified Medical Language System (UMLS)²⁾の医学オントロジーを活用して、文書内の単語の概念に着目することで、類義語や同義語の同定や単語タイプの識別により、目的志向の重み付け手法が開発され、検索や分類などに用いられている³⁻⁶⁾。

医療文書内の医学単語は、例えば hypertension や high blood pressure, hypertonia, ht, hbp のように類義語や同義語および略語による表現の仕方が多く、疾病、治療、薬などの種類も多いため、このような医療文書の特性に対応しつつ、自然言語処理技術により単語の重み付けという前処理をすることは、医療文書分析において本質的であり、重要な課題である。この前処理のプロセスを支援する方法論があれば、膨大な医療情報の二次利用を大きく促進できる。電子カルテの二次利用の整備が進み、利用可能な医学知識も増えている現代において、目的志向の重み付け手法の開発における前処理を支援する汎用性のある方法論を新たに開発することは、医療の分野におけるビッグデータ分析による知識獲得支援の促進として価値があると考えられる。

上記の方法論を開発するにあたり、キーアイデアは、医療文書内の単語を分析目的に応じて、階層的に分類することである。そして、階層的な分類により、複数の単語集合が構成されるが、目的志向の重み付けのため、同じ集合に属する単語には同じ重みを付与し、深い階層に位置する具体的な集合に属すほど、その単語の重みを高くする。ここでの課題は、階

層構造上で、医療文書内の単語を分類する処理をいかに一般化しながらも、分析目的に依存性であるか、という両極端の性質を共存させることであると考えられる。医療分野における単語の分類階層としては、ICD-10⁷⁾が国際的に広く用いられている。ICD-10 の分類目的は診断に特化しているが、統計分類として国際的に広く用いられているため、統計と親和性の高いランキングといった医学関係情報を活用しやすく、分析目的に適した医学関係情報へのマッピングがしやすいと考えられる。

先行研究⁸⁾では、ICD-10 をもとに、死因ランキングという患者リスクの分析に適した医学関係情報を活用して、医学単語集合を分析目的独立のものから依存性の高いものへと順に配置した単語階層を用いて、分析目的限局性の重み付けをすることが、分析精度の向上に有効であることを実験的に示した。特に階層の上位から順に、重み付けに用いる単語の種類を増やすと、分析目的独立の単語集合と目的依存の単語集合の境界で、分析精度の向上が認められた。このことから、目的独立の単語集合の下位に目的依存の単語集合を配置する汎用手法を構成できることが示唆された。

2. 目的

本研究の目的は、医療文書分析の目的限局性に基づいた汎用的な単語階層を構成する方法を検討することである。課題として着目する点は、ICD-10 のコード割り当て対象が疾病であり、医療文書内には疾病以外の薬や検査などの情報も多く含まれるため、ICD-10 のみでは医療文書の分析に不十分なことである。本研究では、この課題に対して、機械学習により近接関係による類似性を活用することを提案する。

医学知識基盤が充実し、機械学習技術が進歩している中、目的志向の重み付けの観点から、既存の医学知識とランキングの医学関係情報を組み合わせ、さらに機械学習を活用することにより、汎用性を保持した分析目的による階層的な単語

の分類体系を作ること、医療ビッグデータ分析による知識獲得支援の促進が期待できると考えられる。

3. 方法

本研究で構成する目的限局性に基づく単語階層は、医療文書を入力として構成され、構成要素は単語集合である。ここで、単語集合はある視点から似た単語の集まりを意味する。単語階層は、階層の上位に位置する医療文書分析の目的と独立する部分と、階層の下位に位置する目的に依存する部分の2つからなる。目的独立と目的依存部分の単語階層に配置される単語集合は、どちらも医療文書内の単語集合の部分集合となる。つまり、単語集合内の単語は、医療文書に出現する単語のことを意味し、ICD-10などの医学知識の中の単語ではない。ここで、単語とは、ある1語だけではなく、複数の語で構成される用語も含む。分析目的と独立する部分と、分析目的に依存する部分を構成するための方法をそれぞれ3.1と3.2で述べる。

3.1 分析目的と独立する部分の構築

医療文書の分析目的に応じて、重み付けの際に重要となる単語は変化すると考えられる。しかしながら、同じ単語集合に属すれば同じ重みを与え、階層が深くなることにより具体的な単語集合で構成される単語階層であれば、その上位の階層は医療文書のどの分析目的にも共通して重み付けに活用できると考えられる。そこで、本研究では単語階層において、目的に依存する部分を切り離して、分析目的とは独立した部分を構築する。その構築方法を順に述べる。

まず、医療文書を入力として、医療文書から医学単語を識別する。そのため、医学単語の概念を扱うことができるUMLSの医学オントロジーを用いる。UMLSによるコードで示された医学単語の概念情報を用いることで、同義語や類義語の問題に対応できる。具体的には、MetaMap⁹⁾により、医療文書内の単語のUMLSの概念コードを取得する。そして、概念コードを持つ医療文書内の単語を医学単語集合へ、そうでない単語を一般単語集合へ分類する。

次に、医学単語集合をさらに、ICD-10医学単語集合と非ICD-10医学単語集合へ分類する。ICD-10は、国際的に広く統計分類として用いられていることから、ICD-10は統計と親和性の高いランキングの医学関係情報を活用しやすく、分析目的に適した医学関係情報へのマッピングに有効であるため、分析目的に依存する部分への連結に適していると考えられる。

医療文書内の医学単語のICD-10コードを取得するために、UMLSの概念コードとBioPortal¹⁰⁾を用いる。ICD-10コードを持つ医学単語をICD-10医学単語集合へ、そうでない医学単語を非ICD-10医学単語集合へ分類する。

上記のように医学知識を組み合わせることで、医療文書から、分析目的と独立する部分を構築する。

3.2 分析目的に依存する部分の構築

3.1により、分析目的に依存する部分を分離させることにより構築した分析目的と独立する部分を固定した上で、その下位に目的依存の単語階層を構築する。本研究では、ICD-10を活用して医学単語が持つ様々な特性を捉えることで、分析目的に依存する単語階層を、設定する分析目的に応じて変化させることが可能になる。つまり、目的独立部分と目的依存部分の橋渡し役として用いるのが、診断に関する分析目的に適したランキングの医学関係情報へのマッピングに有効な

ICD-10である。例えばICD-10の中には、hypertensive heart diseaseというラベル(ICD-10単語)があり、I11のICD-10コードに対応している。このICD-10単語が、分析目的に応じて様々な特性を持つことを、例を挙げながら述べる。例えば、患者リスクを予測する分析や、在院日数を予測する分析において、それらの分析目的に依存する単語階層を構築するために、ICD-10コードを活用して、死因ランキング¹¹⁾や平均在院日数¹²⁾の医学関係情報にICD-10単語をマッピングできる。なぜなら、それらの医学関係情報はICD-10に紐づけられているためである。例えば、死因ランキングの場合、I11はランク1に該当し、平均在院日数の場合、I11は8.3日に該当する。そのため、患者リスクの予測という分析目的の観点からは、hypertensive heart diseaseというICD-10単語は、死因ランキング1位に該当することから、患者の重症度合いが高いという特性を持つ。平均在院日数の予測という分析目的の観点からは、上記ICD-10単語は、平均在院日数で8.3日に該当することから、長期入院の度合いが低いという特性を持つ。これらの特性は、患者リスクの予測と在院日数の予測というそれぞれの分析目的に適しており、重み付けにおいても重要な要素であると考えられる。このように、ICD-10単語は、診断に関するランキングの医学関係情報にマッピングさせることで、扱う分析目的に応じて様々な特性を持つことがわかる。そこで、本研究では、分析目的に適したICD-10対応のランキングへマッピングできるICD-10医学単語集合内の単語を、分析目的に依存するICD-10医学単語集合へ分類し、そうでない場合は分析目的に依存しないICD-10医学単語集合へ分類する。分析目的に依存するICD-10医学単語集合内の単語は、該当するランキング内のランクごとでさらに分類がされる。

本研究ではさらに、ある分析目的に適したランキングの医学関係情報がカバーできないランキング外の医療文書内の単語についても、近接関係により類似するランキング内のICD-10医学単語の重みを参考に、分析目的に応じて重みを可変することが、対象となる分析目的における分析精度の向上に寄与すると考えられる。例えば、ST dep(ST低下)という医学単語は、心電図による異常な所見であり、心疾患の診断において重要な情報であるが、ICD-10コードに対応しないため見過ごされてしまう。そこで、シソーラスでは表されない単語の近接関係に基づいて、一般単語集合と非ICD-10医学単語集合および分析目的に依存しないICD-10医学単語集合の下位に、分析目的に類似的に依存する単語集合を新たに構築する。ここで、類似とは、単語の共起に基づいて学習される類似性を指し、Word2Vec¹³⁾により単語をベクトルに変換した上で、コサイン類似度により単語間の類似性を求める。分析目的に依存するICD-10医学単語集合内のそれぞれの単語と、上記の3つの単語集合内の単語との類似度を計算し、閾値以上の類似度を持つ上記の3つの単語集合内の単語を、それぞれの3つの単語集合の下位に配置される、分析目的に類似的に依存する単語集合へ分類し、そうでない場合は分析目的に類似的に依存しない単語集合へ分類する。分析目的に類似的に依存する単語集合内の単語は、該当するランキング内のランクごとでさらに分類がされる。

上記のように、ランキングの医学関係情報と機械学習を活用して、分析目的に依存する部分を構築する。

4. 結果

本研究で検討した目的限局性に基づく単語階層は図1となる。図1内の上の赤枠が分析目的と独立する部分で、下の赤枠が分析目的に依存する部分である。

分析目的と独立する部分の単語階層は、UMLS と ICD-10 の活用により、最上位で一般単語集合と医学単語集合で構成され、医学単語集合は ICD-10 医学単語集合と非 ICD-10 医学単語集合から構成された。分析目的に依存する部分の単語階層は、ICD-10 とランキングの医学関係情報の活用により、ICD-10 医学単語集合の下位に、分析目的に依存する・依存しない ICD-10 医学単語集合という 2 つの単語集合が構成された。また、ランキングの医学関係情報と機械学習の活用により、単語の共起により学習される類似性に基づいて、一般単語集合と非 ICD-10 医学単語集合および分析目的に依存しない ICD-10 医学単語集合の下位に、分析目的に類似的に依存する単語集合が構成できることが示唆された。分析目的に依存する単語集合および分析目的に類似的に依存する単語集合の下位には、ランキング内のランクごとに該当する単語集合が構成され、図 1 では、ランク 1 に該当する単語集合とランク 2 に該当する単語集合およびランク n に該当する単語集合と表記している。ここで、 n はランキング内のランク数である。

5. 考察

提案手法は、ランキングの医学関係情報だけでなく、近接関係による類似性を求められる機械学習の Word2Vec を活用することで、ICD-10 が対象とする疾病に関する単語だけでなく、一般単語や非 ICD-10 医学単語など様々な種類の単語を分析目的に類似的に依存する単語集合の要素として、単語階層の分析目的に依存する部分の対象とすることを可能にした。したがって、提案手法は、ICD-10 の対象範囲の課題に対処できると考えられる。また、分析目的に依存する疾病に関する単語と、分析目的に類似的に依存する疾病や薬、検査などに関する医療文書内の様々な種類の単語との関係性について、該当する類似度を用いながら分析できるようになることは、知識発見の観点から意義があると考えられる。さらに、ランキングの医学関係情報の活用により、分析目的軸から ICD-10 単語の様々な特性を捉えられるようになることで、医療文書内の単語の意味を考慮した分析に役立つと考えられる。

しかしながら、例えば、ICD-10 単語の hypertensive disease は、I10-I15 の ICD-10 コードに対応し、そのコードは I10, I11, I12, I13, I14, I15 の 5 つのコードに該当することになる。患者リスクの予測という分析目的に適した死因ランキングを参照すると、I10 と I12, I15 のコードは死因ランキングのランク 13 に該当し、I11 と I13 のコードは死因ランキングのランク 1 に該当する。このように、あるランキングの中で複数のランクに該当するような粒度の粗いコードに対応する単語の分類方法と、それに伴う重み付け方法を考える必要がある。また、類似度計算において、例えば、ある分析目的に依存する ICD-10 医学単語集合の下位に、ランクごとで 3 つの単語集合がある場合において、3 つの単語集合すべてに対して閾値以上の類似度を持つ単語の扱い方を検討する。

提案する単語階層を重み付けに適用するには、下記の 4 つの性質を活用できると考えられる。1 つ目は、深い階層に位置する単語集合に対してより高い重みを与えることである。例えば、一般単語集合よりも非 ICD-10 医学単語集合に高い重みを与えられる。2 つ目は、論理否定の活用により、A である場合に A でない場合の単語集合よりも高い重みを与えることである。例えば、非 ICD-10 医学単語集合よりも ICD-10 医学単語集合に高い重みを与えられる。3 つ目は、ランキングの医学関係情報を活用して、高いランクに該当する単語集合であ

ればあるほど、高い重みを与えることである。例えば、ランキング内でランク 1 に該当する単語集合が他のランクに該当する単語集合の中で最も高い重みを与えられる。4 つ目は、近接関係に基づく類似性により、分析目的に依存する単語集合の下位に位置する、ランキング内のランクごとに該当する単語集合に与えられる重みを、ランクに該当する単語集合に類似的に依存する単語集合内の重みに反映させることである。例えば、非 ICD-10 医学単語であっても、ランキング内のランク 1 に該当する単語集合に類似的に依存している場合は、ランク 1 と同程度の重みを与えられる。

6. 結論

本研究は、医療文書分析における重み付けという前処理を支援する方法論として、既存の医学知識の UMLS と ICD-10 および ICD-10 に対応する診断に関するランキングの医学関係情報の組み合わせと、機械学習の活用により、汎用性を保持した目的限局性に基づく単語階層の構成法を検討した。提案する単語階層は単語集合で構成され、分析目的と独立する部分と分析目的に依存する部分の 2 つからなることとした。特に課題として、ICD-10 が疾病に関する単語を対象としているのに対し、医療文書内の単語には、疾病以外の、薬や検査などに関する単語も多く含まれているため、医療文書分析には ICD-10 のみでは不十分であることとした。そこで、本研究では、近接関係に基づく類似性を捉えられる機械学習の Word2Vec を活用することで、ランキング対象外の ICD-10 医学単語および ICD-10 が対象とする疾病以外の単語においても、分析目的に類似的に依存する単語集合へ属することを可能にした。提案する単語階層は、同じ単語集合に属する単語には同じ重みを付与するという基本的な性質から、階層関係と論理否定の関係、ランキングの関係および類似性の関係という 4 つの関係性の性質を活用することで、目的志向の重み付けに役立てられると考えられる。今後は、考察で述べた課題を検討しながら、階層的な単語分類体系を作ることで、医療ビッグデータ分析による知識獲得支援の促進が期待できる。

参考文献

- 1) The National Library of Medicine (NLM). Medical Subject Headings (MeSH). [https://www.nlm.nih.gov/mesh/ (cited 2017-Sep-8)].
- 2) Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 2004;32(suppl 1):D267–D270.
- 3) Zhang X, Jing L, Hu X, Ng M, Zhou X. A comparative study of ontology based term similarity measures on PubMed document clustering. In: *International Conference on Database Systems for Advanced Applications* 2007;115–126.
- 4) Zhang X, Jing L, Hu X, Ng M, Jiangxi JX, Zhou X. Medical document clustering using ontology-based term similarity measures. *International Journal of Data Warehousing and Mining* 2008;4(1):62–73.
- 5) Yu H, Cao YG. Using the weighted keyword models to improve information retrieval for answering biomedical questions. In: *Summit on Translational Bioinformatics* 2009;143–147.
- 6) Zhu W, Xu X, Hu X, Song IY, Allen RB. Using UMLS-based re-weighting terms as a query expansion strategy. In: *IEEE International Conference on Granular Computing* 2006;217–222.
- 7) World Health Organization. International statistical classification of diseases and related health problems. 2004;1.
- 8) 松尾亮輔, Ho Tu Bao. 医学文書分析における意味を考慮した単

語重み付けのための医学知識表現. 第7回知識共創フォーラム
2017.

- 9) Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *In: AMIA Symposium* 2001;17–21.
- 10) Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* 2009;37:W170–W173.
- 11) Murphy SL, Xu J, Kochanek KD. Deaths: final data for 2010. *NVSR* 2013;61(4):1-118.
- 12) Kramers, PG. The ECHI project: health indicators for the European Community. *The European Journal of Public Health* 2003;13:101–106.
- 13) Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. *In: Advances in Neural Information Processing Systems* 2013;3111-3119.

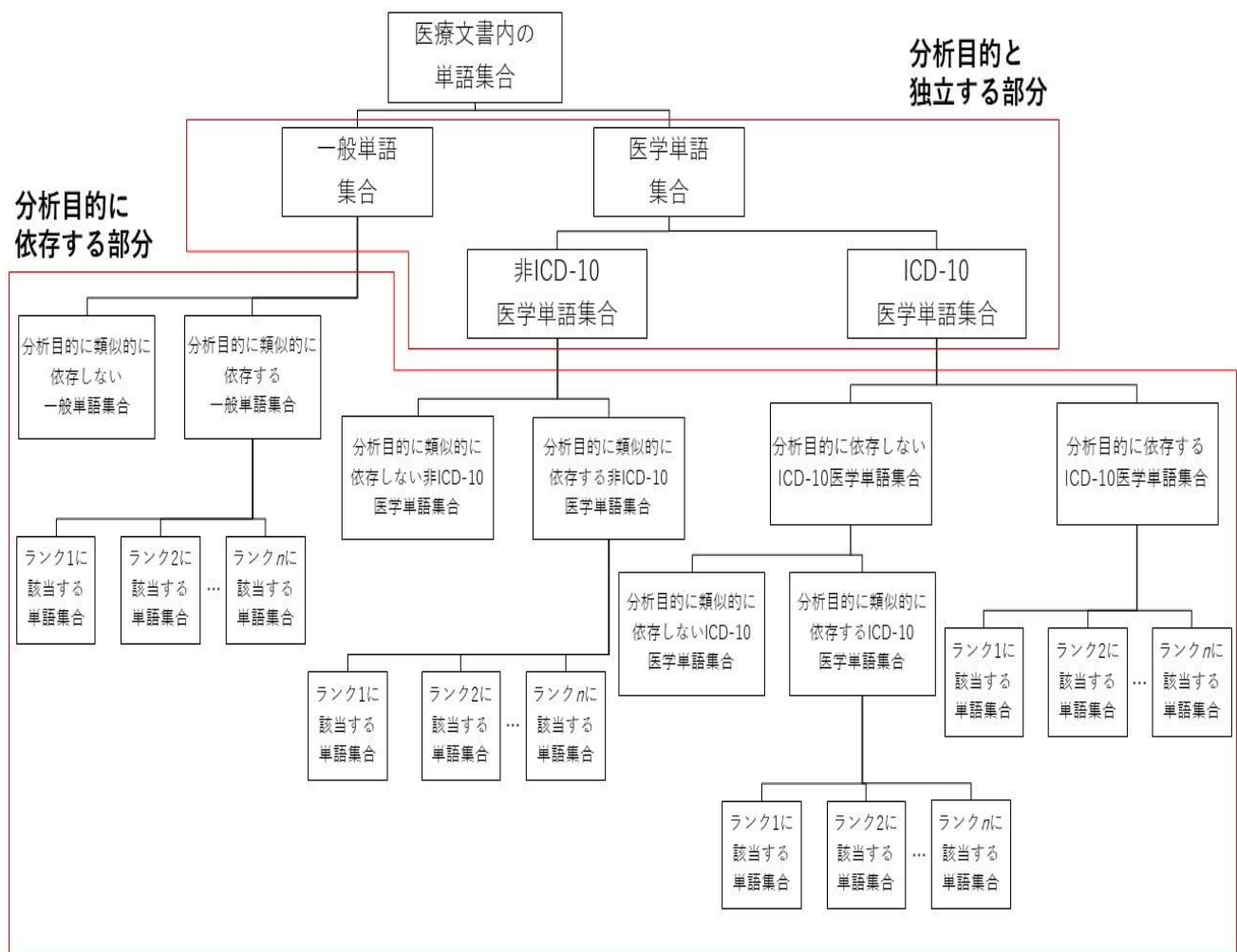


図 1 医療文書分析の目的限局性に基づく単語階層