

QA サイトでの共起に基づく患者の自覚症状入力支援

新本 拓也¹ 湯本 高行¹ 金子 周司² 磯川 悌次郎¹ 松井 伸之¹ 上浦 尚武¹

概要: 本研究では、ユーザが問診などで症状や症状が起こった部位などを入力する際にそれを支援するための QA サイトを利用した語彙の推薦手法を提案する。QA サイトの質問テキストから抽出した症状とその部位を利用して相関ルールとして表される共起関係を収集する。得られた相関ルールに対して、入力された症状や部位のクエリと対応する条件部を持つ相関ルールの帰結部である症状と部位をランキングして、推薦結果として出力する。この推薦結果は入力に対して同時に発生するような症状や入力の言い換え表現が含まれ、自覚症状として問診などの際の入力を促すことができる。評価として 10 のクエリに対して提案手法を適用したところ、症状「痛み」+部位「目」に対して症状「頭痛」、症状「頭痛」に対して症状「痛み」+部位「頭」などを推薦できた。

1. はじめに

医療現場において医療従事者の医療以外の事務作業などに費やす時間を削減し、より患者や医療に向き合える体制を構築していくことは今後の医療の発展に必要な不可欠な課題である。しかし、医師の勤務時間の割り当てにおいて多くの時間が電子カルテの入力などの事務作業に多くの時間が費やされているという報告も存在する [1]。ここで問診などでの自覚症状の入力を患者自身が行えるようになることができれば待合などでの時間で入力を行い、医療従事者の事務作業の時間を削減できるのではないかと考えている。

また、一つの課題として自覚症状の入力を行う際に、専門用語は患者などの一般のユーザには分からないという問題がある。そこで一般ユーザにも理解がしやすく、多様性のある電子データとして QA サイトでの質問記事を利用する。これにより、よりユーザに身近な表現を用いた入力支援が行える可能性がある。また、専門家でないユーザに自覚症状の入力を求めたときに、それに関連する症状を自覚していても申告しない可能性があるということも問題として挙げられる。

本研究では、ユーザが問診などで症状や症状が起こった部位などを入力する際にそれを支援するための QA サイトを利用した語彙の推薦手法を提案する。QA サイトの質問テキストから抽出した症状とその部位を利用して相関ルールとして表される共起関係を収集する。得られた相関ルールに対して、入力された症状や部位のクエリと対応する条件部を持つ相関ルールの帰結部である症状と部位をランキ

ングして、推薦結果として出力する。この推薦結果は入力に対して同時に発生するような症状や入力の言い換え表現が含まれ、自覚症状として問診などの際の入力を促すことができる。

たとえば症状として「咳」と部位として「喉」が入力されたときに、関連する症状として「痛み+のど」を出力し、関連する自覚症状としての入力を促すことなどができる。

2. 関連研究

2.1 使用する言語リソース

本研究では言語リソースとして MEDIS の ICD-10 対応標準病名マスター [2] やライフサイエンス辞書 (LSD) [3] を使用している。

ICD-10 とは世界保健機関 (WHO) が作成した疾病や死亡のデータの体系的な記録、分析、解釈及び比較を行うための分類である [4]。ICD-10 対応標準病名マスターはこの ICD-10 に基づいて標準的な日本語の病名と病名に対応する ICD コードを収録した基本マスターとなっている。

また、ICD-10 標準病名マスターだけでは表記揺れに対応できないため、LSD も併用する。LSD は生命科学分野の専門用語に関する類語辞典で、頻繁な用語の追加や改訂で先進的な用語も取り入れられている。さらにこの LSD には Medical Subject Heading (MeSH) に対応した用語の階層構造が概念ツリーとして収録されている。MeSH はアメリカ国立医学図書館の制定した文書の索引付け、目録作成、検索に使用される用語集であり、主要カテゴリから階層構造を下位に移していくことで厳密な定義語を得ることができる [5]。たとえば MeSH Tree Structure では

¹ 兵庫県立大学大学院工学研究科

² 京都大学大学院薬学研究科

Extremities(UniqueID : D005121) の下に Lower Extremity(D035002), Upper Extremity(D034941), Amputation Stumps(D000672) が与えられる。これは LSD においては四肢 (Extremities) の下に下肢 (Lower Extremity), 上肢 (Upper Extremity), 切断端 (Amputation Stumps) としてツリー構造ごと日本語訳されて収録されている。本研究では LSD に収録されたこの概念ツリーを利用し、表記揺れの統一などを行う。

2.2 相関ルール

相関ルールとは A と B を個々のアイテムとして含むトランザクションが与えられたとき、A を含むトランザクションが B を含むことを示すルールである [6]。相関ルールは一般に $A \Rightarrow B$ で表され、A を相関ルールにおける条件部、B を帰結部という。相関ルールの評価指標として支持度 (support), 確信度 (confidence), リフト値 (lift) が挙げられる。支持度は A と B の共起確率であり、確信度は A が与えられたときに B が出現する条件付き確率、 $A \Rightarrow B$ のリフト値は確信度を B の支持度で割ったものとなっている。相関ルールの探索はある A や B をアイテムとして含むトランザクションのセットにおいてユーザの指定した最小支持度、最小確信度を超える全てのルールを発見する問題で、これによって得られたルールは同時性と関連性が高いルールと期待できる。

相関ルールの探索は一般的に Apriori アルゴリズムを用いられる [7]。このアルゴリズムは通常、より一般化されたアイテムの表現や上位下位概念を考慮したものにはなっていない [8]。たとえば、衣服-アウター-ジャケットや履き物-靴といった分類構造を持つアイテムの相関ルールは、通常「靴」と「ジャケット」といった最も下層のアイテム同士でのルール抽出が想定される。そのため、「アウター \Rightarrow 靴」のようなルールはそのままでは抽出されない。これの解決には、用いるトランザクションへのアイテムの追加や、分類木構造を利用したルール抽出の効率化などの工夫が必要となる。本研究では、部位の表現に LSD に収録された部位の概念ツリーを用いるため、このようなツリーの構造上上位に存在する部位表現との相関ルール抽出が必要とされる。

3. 提案手法

本研究では QA サイトの質問記事から相関ルールを抽出し、そのルールを用いて推薦を行う。推薦では入力症状や部位に対して、言い換えの表現や同時に発生しやすい症状を出力として得ることができ、これによって、ユーザに対して入力症状と同時に出力のような症状が発生していないか問いかけ、自覚症状として把握させることができる。なお、一連の流れを図 1 に示す。本研究では情報システム研究機構国立情報学研究所の情報学研究データリポジトリから提供を受けた Yahoo! データセット [9] の「病気、症状、

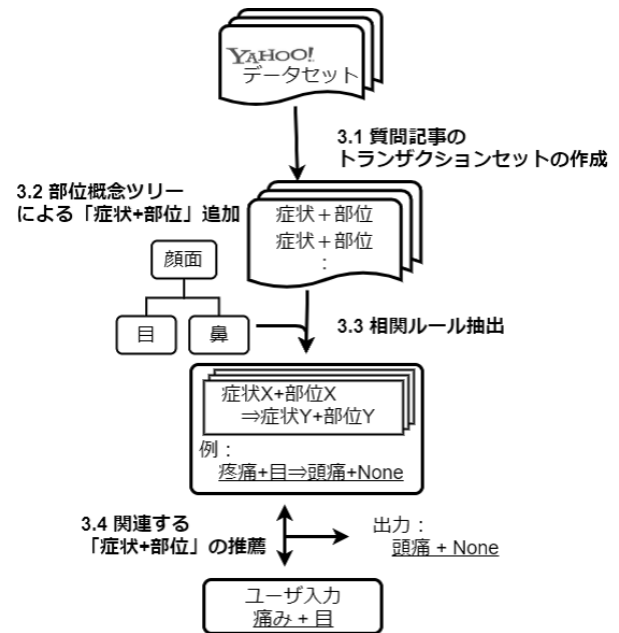


図 1 提案手法の概要の一部

ヘルスケア」カテゴリ質問記事 75000 件を使用する。

3.1 質問記事からの症状トランザクションセットの作成

推薦に用いるルール抽出のため、Yahoo! データセットの質問記事からトランザクションのセットを作成する。各トランザクションは症状と起こった部位をセットにしたアイテム (以下「症状+部位」) を含んでおり、トランザクション中での「症状+部位」の共起から相関ルールを抽出する。そのため、各トランザクションとなる質問記事から症状と部位を抽出する必要がある。この症状と部位の抽出は QA サイトから病訴検索する手法 [10] の一部を基に行う。

3.1.1 症状の抽出

症状の抽出は質問テキストの形態素解析で得られた形態素と、病名マスターの形態素解析で得られた形態素の比較によって行う。形態素に分解した各質問記事のそれぞれのテキストについて、病名マスター中の病名やその表記揺れとなる病名とに共通する名詞数から症状の候補を決定し、形態素単位での距離から候補とテキスト中の対応する部分の同定を行う。たとえば、「軽い腹痛や腰痛があります」というテキストと、「腰痛症」という病名があるなら、テキストと病名の共通の名詞数は「腰痛」の 1 つだけであり、これに対して病名の名詞数は 2 であるため、スコアとして $1/2$ が与えられる。このスコアが上位のものをそのテキストで出現する症状の候補として決定し、その候補とテキストの対応部分を形態素単位の距離を比較する。「軽い腹痛や腰痛があります」というテキストであれば、候補となった「腰痛症」ともっとも近い距離を持っているのが下線の部分であるため、そこがテキストと症状の対応部分であると決める。このとき、得られた症状は ICD-10 コード付き

の症状として保存される。

3.1.2 関連部位の抽出

関連部位は係り受け解析からルールベースの手法で行う。係り受け解析には JUMAN 辞書 [11] を適用した CaboCha [12] を利用する。JUMAN 辞書には解析時に各形態素にその形態素のカテゴリ情報が付与される。それによって解析された形態素に「カテゴリ：動物-部位」を含むものを部位と判定し、その形態素が含まれる文節を抽出する。さらに、LSD に含まれる部位の類似表現を含む文節も部位と判定する。また、ルールの抽出に「症状+部位」でセットにしているため、抽出された症状を含む文節と部位表現を含む文節の係り受け関係によって症状と部位を対応付ける。この際、係り受けによって対応する部位が存在しない症状もアイテムとして認める。

3.2 部位の概念ツリーの構築

抽出を行った部位は表記揺れが多く含まれている。そのため、この表記揺れを統一する手段として LSD に収録されている部位の概念ツリーを利用した部位表現の拡張を行う。得られた部位表現を LSD の部位概念ツリーに対応付けることで、類義語や上位下位概念を扱うことができるようになる。部位表現と LSD の概念ツリーとの対応付けは以下の三段階によって行う。

- (1) LSD の類語関係を使用した完全一致
- (2) Wikipedia のリダイレクト・リンクアンカー関係を用いた完全一致
- (3) 既に関連づけられた部位情報との部分一致

(1) は LSD に含まれる類語辞書を用い、概念ツリーの各ノードに対象の部位表現を対応させる段階である。(2) は対象の部位表現を Wikipedia のリダイレクト・リンクアンカー関係によって概念ツリーの各ノードに対応付ける段階となっている。リダイレクトとは Wikipedia で略称や別表記を検索した際にその語に対応する実際の記事へと転送する機能であり、リンクアンカーは Wikipedia の別の記事の本文中の語をその語の記事にリンクさせる機能である。どちらも異表記を一つの記事へと集約する機能であり、この関係を用いることによって LSD の類語辞書では弱かった部位表現の日常的な表記を部位の概念ツリーに対応づけることができる。たとえば、「まぶた」という単語は LSD の概念ツリーでは登録されていない単語であるが、リダイレクト関係からは「目蓋、瞼、二重まぶた、眼瞼」といった言葉がリダイレクト元として得られ、リンクアンカー関係からは「一重まぶた、二重瞼、眼瞼」などの単語が得られる。このうち、「眼瞼」という単語が LSD の概念ツリーに含まれているため、これによって「眼瞼」から「まぶた、瞼、目蓋」などの表記揺れを概念ツリーに対応づけられる。さらに (3) により、たとえば「手」と「右手」など部分一致で結びつけられるような部位情報を概念ツリーに対応づ

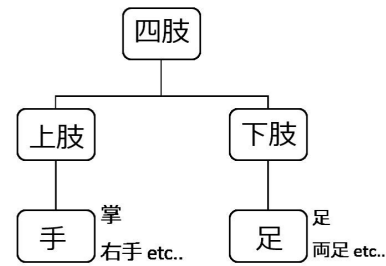


図 2 部位概念ツリーの一部

ける。得られた概念ツリーの一部を図 2 に示す。

3.3 相関ルールの抽出

本研究では、3.1 の手法で作成したトランザクションに部位を抽象化した「症状+部位」を追加した後に相関ルールの検出を行う。具体的には以下のステップで行う。

- (1) 「症状+部位」の部位表現の置換
- (2) 部位の抽象化による「症状+部位」の追加
- (3) パスの前方一致によるルール抽出時の選別

(1) ではトランザクションの「症状+部位」について、拡張した部位の概念ツリーを用いて部位表現を概念ツリーの代表表記で統一し、表記揺れにより頻度が低下するのを防ぐ。たとえば「痛み+掌」というような「症状+部位」がトランザクションに含まれているとすると、これを「痛み+手」へと置換する。これによって「痛み+右手」などの「症状+部位」があったときに、「痛み+手」に置換されることで表記揺れの統一がされ、最小支持度を超えやすくなる。

(2) ではトランザクションに含まれる「症状+部位」の部位表現を部位の概念ツリーに基づく上位カテゴリに置換した新たな「症状+部位」をトランザクションに追加する。たとえば「痛み+手」といった「症状+部位」があったとすれば、部位の概念ツリーに基づくとその上位カテゴリは「上肢」「四肢」などになるため、トランザクションに「痛み+上肢」「痛み+四肢」といった新たな「症状+部位」を追加する。

相関ルールの抽出には、最小支持度を 30、リフト値を 1 以上に設定し、Apriori アルゴリズムを用いる。確信度は「症状*部位」を推薦する際に利用するが、ここでは用いない。抽出の際、(2) のトランザクションの拡張によって自明なルールが発生する。たとえば「痛み+手」⇒「痛み+上肢」といったルールは同じ「症状+部位」から (2) の追加によって発生した「症状+部位」同士であるため、ルールとするには明らかに不適当である。そこで (3) では、部位の概念ツリーのルートからのパスの前方一致で合致する部位同士のルールを不適当なものとして除外する。同じ症状で「症状+部位」の部位表現が「手」と「上肢」ならば、部位の概念ツリーでのパスはそれぞれ「身体領域/四肢/上肢/手」と「身体領域/四肢/上肢」になるため、後者

が前者と前方一致する。したがってこの組み合わせの時はルールとして抽出を行わないようにする。

以上の手法によって Yahoo!データセットの質問記事から相関ルールとして関連する「症状+部位」のペアを抽出し、推薦に用いる。

3.4 関連する「症状+部位」の推薦

ユーザが入力した「症状+部位」に対して関連する「症状+部位」を推薦する。具体的には、ユーザが入力した「症状+部位」(Z_symp+Z_part)に対応する条件部(X_symp+X_part)を持つ相関ルールの結論部(Y_symp+Y_part)を推薦する。詳細は以下のとおりである。

- (1) Z_symp に対応する X_symp を条件部に含む相関ルールの発見
- (2) (1) で得られた相関ルールの X_part と Z_part を比較
- (3) (2) の結果によってスコアを変え、(Y_symp+Y_part) をランキング

(1) において入力した症状クエリ (Z_symp) は編集距離を利用して類似する病名の ICD10 コードに変換し、相関ルールを検索する際の入力として使用する。事前に得られた相関ルールの X_symp が ICD10 コードに変換した症状クエリ (Z_symp) と一致するルールを取り出しておく。

入力された部位クエリ (Z_part) は部位の概念ツリーによる代表表記化を行っている。また、類似する文脈で出現する部位表現を得るため、Yahoo!データセットを学習させた fastText[13] による単語分散表現のコサイン類似度を利用する。ここで類似表現を得る際は、コサイン類似度が 0.65 以上のものを採用している。得られた部位クエリの集合を Z'_part とする。(1) で得られたルール群の X_part が Z'_part と対応するものとし、ないもので二つに分ける。

相関ルールの帰結部のうち部位が対応したもの (A) は与えられた確信度をそのままスコアとして利用し、対応しなかったルール (B) は与えられた確信度に A の最大の確信度の半分をかけたものをスコアとして与える。A と B を合わせてスコア順にランキングし (Y_symp+Y_part) を最大 10 件まで推薦結果として出力する。B を結果に含めるのは、たとえば「アレルギー」という症状クエリに「鼻」といった部位クエリで入力した際に他のアレルギーの症状なども同時に発生するような推薦の結果として含めるためである。

推薦の流れを図 3 に示す。

4. 評価実験

評価のための実験として、一般的に入力されやすいであろう表 1 の 10 のクエリを用意し、実際に推薦を行うことによって関連する「症状+部位」が推薦されているかを確認した。正解となる結果は入力クエリに関連する症状や言い換え表現となっているものと定義し、著者のうち 1 名が判定を行った。表 2 から表 11 に示す推薦の結果には初め

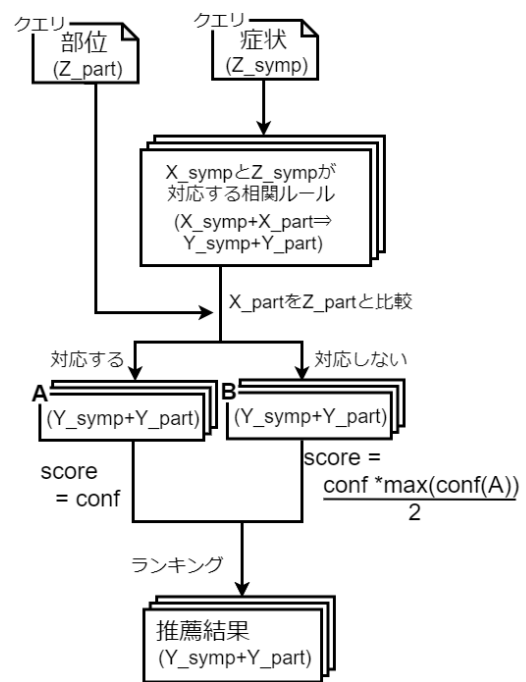


図 3 関連する「症状+部位」の推薦

表 1 クエリとその評価

| 病名 | 部位 | RR |
|---------|----|------|
| 痛み | 足 | 0.14 |
| 痛み | 目 | 0.50 |
| 腫れ | のど | 0.33 |
| 風邪 | | 1.0 |
| インフルエンザ | | 0.33 |
| 吐き気 | | 1.0 |
| 咳 | 喉 | 0.50 |
| 頭痛 | | 1.0 |
| 腹痛 | | 1.0 |
| アレルギー | | 1.0 |
| 平均 | | 0.68 |

て出てきた正解に○を付けている。これにより推薦された結果からそれぞれの逆順位 (RR) とその平均逆順位 (MRR) を調べた。

これによって一般的に検索されることが多いと考えられる 10 のクエリに対しては約 2 位前後で有効な「症状+部位」が推薦されることが分かった。クエリ「アレルギー」では、花粉症や喘息、鼻炎などの関連する「症状+部位」が推薦結果の上位に入った。また、「痛み+目」の推薦で現れた「頭痛」は、群発性頭痛の症状の一つであり、「吐き気」の推薦で現れた「頭痛」も片頭痛などで現れる症状の一つであるため、類似の症状を推薦できていると評価できる。問題として「痛み」や「腫れ」といった抽象的なクエリに対して図 3 の B の評価を適用した結果、推薦される「症状+部位」に不適当なものが混ざることが挙げられる。たとえば「腫れ+のど」に対して得られた「腫脹+咽頭」といった結果は、入力も出力も同じものであるため部位もマッチ

表 2 痛み+足

| 正 | 症状 | 部位 |
|---|-----|-------|
| | 疼痛 | 上肢 |
| | 疼痛 | 手 |
| | 疼痛 | 手指 |
| | 疼痛 | 骨格 |
| | 腫脹 | 四肢 |
| | 腫脹 | 下肢 |
| ○ | 腫脹 | 足 |
| | 疼痛 | 生殖器 |
| | しびれ | 四肢 |
| | 神経症 | ニューロン |

表 3 痛み+目

| 正 | 症状 | 部位 |
|---|----|----|
| | 疼痛 | 四肢 |
| ○ | 頭痛 | |
| | 頭痛 | 頭部 |
| | 頭痛 | 顔面 |
| | 頭痛 | 眼 |
| | 腫脹 | 頭部 |
| | 腫脹 | 顔面 |
| | 腫脹 | 眼 |
| | 眼振 | 眼 |
| | 眼振 | 頭部 |

表 4 腫れ+のど

| 正 | 症状 | 部位 |
|---|----|-------|
| | 腫脹 | リンパ系 |
| | 腫脹 | 免疫系 |
| ○ | 腫脹 | リンパ組織 |
| | 疼痛 | 咽頭 |
| | 疼痛 | 消化管 |
| | 腫脹 | 骨格 |
| | 腫脹 | 咽頭 |
| | 腫脹 | 中咽頭 |
| | 腫脹 | 女性生殖器 |
| | 腫脹 | 子宮付属器 |

表 5 風邪

| 正 | 症状 | 部位 |
|---|---------|-----|
| ○ | 咳 | |
| | 疼痛 | 消化管 |
| | 疼痛 | 咽頭 |
| | かぜ | 咽頭 |
| | かぜ | 中咽頭 |
| | 頭痛 | |
| | 高熱 | |
| | 喀痰 | 体分泌 |
| | インフルエンザ | |
| | 喀痰 | 痰 |

表 6 インフルエンザ

| 正 | 症状 | 部位 |
|---|------|-----|
| | 予防接種 | |
| | かぜ | |
| ○ | 高熱 | |
| | 咳 | |
| | 頭痛 | |
| | 疼痛 | 消化管 |
| | 疼痛 | 咽頭 |

表 7 吐き気

| 正 | 症状 | 部位 |
|---|-----|-------|
| ○ | 腸炎 | |
| | 頭痛 | |
| | めまい | |
| | 疼痛 | 胴体 |
| | 疼痛 | 腹部 |
| | 疼痛 | 消化管 |
| | 疼痛 | 上部消化管 |
| | 疼痛 | 胃 |
| | かぜ | |
| | 腹痛症 | |

表 8 咳+喉

| 正 | 症状 | 部位 |
|---|----|-----|
| | 疼痛 | 消化管 |
| ○ | 疼痛 | 咽頭 |
| | かぜ | |
| | 喀痰 | 体分泌 |
| | 喀痰 | 痰 |
| | かぜ | 咽頭 |
| | かぜ | 消化管 |
| | 疼痛 | 胴体 |
| | 咳 | 体分泌 |
| | 咳 | 痰 |

表 9 頭痛

| 正 | 症状 | 部位 |
|---|-----|-----|
| ○ | 嘔気 | |
| | 疼痛 | 頭部 |
| | かぜ | |
| | めまい | |
| | 肩こり | |
| | 疼痛 | 消化管 |
| | 疼痛 | 咽頭 |
| | 疼痛 | 胴体 |
| | 倦怠感 | |
| | 疼痛 | 首 |

表 10 腹痛

| 正 | 症状 | 部位 |
|---|-----|-----|
| ○ | 腸炎 | |
| | 嘔気 | |
| | 疼痛 | 胴体 |
| | 疼痛 | 腹部 |
| | 便秘症 | |
| | 疼痛 | 消化管 |
| | 頭痛 | |
| | かぜ | |

表 11 アレルギー

| 正 | 症状 | 部位 |
|---|------|----|
| ○ | 花粉症 | |
| | かぜ | |
| | 咳 | |
| | 咳喘息 | |
| | 小児湿疹 | |
| | 鼻炎 | |
| | 湿疹 | |
| | 麻疹 | |
| | くしゃみ | |
| | 腫脹 | 頭部 |

したAの「症状+部位」では取り除かれているものである。また、利用した質問記事が比較的少ないことから、多くの入力クエリでは症状と部位が両方一致する相関ルールの条件部が少なく、Bの症状クエリにのみ頼った推薦を行ってしまう問題もある。さらに、「痛み+足」に出てきているような「神経症+ニューロン」といった質問記事からのアイテム抽出段階での間違いも存在していた。

5. おわりに

本研究ではQAサイトの質問記事から症状とそれが起こった部位をセットにした「症状+部位」を抽出し、1つの質問記事を1つのトランザクションとしたデータセットから得られた相関ルールを利用した推薦を行った。相関ルールの抽出ではより幅の広いルールの抽出を行うため、得られたトランザクションに部位の概念ツリーに基づいて部位を抽象化した「症状+部位」を追加した。これによって得られた関連する「症状+部位」の推薦結果を一般的に頻出するであろう10のクエリを用いて評価した結果、多くのクエリでは推薦の上位に有効な「症状+部位」が与えられた。

今後の課題として相関ルール抽出のベースとなるトランザクションの増加や相関ルールの評価指標を変化させることにより、網羅性の高い推薦を行えるようにすることが挙げられる。また、推薦の柔軟性を上げるため、症状や部位といったクエリを直接与えるのではなく、入力されたテキストからクエリを取り出し推薦を行うシステムの開発も検討している。

謝辞 本研究の一部は科研費基盤研究(C)(17K00429)によるものである。

参考文献

- [1] C. Sinsky, L. Colligan, L. Lii, et al: *Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties*. Annals of internal medicine 165.11, pp.753-760, 2016.
- [2] 一般財団法人医療情報システム開発センター (MEDIS-DC): ICD10 対応標準病名マスター, <http://www2.medis.or.jp/stdcd/byomei/index.html>
- [3] LSD プロジェクト: Life Science Dictionary, <https://lsd-project.jp/ja/index.html>
- [4] 厚生労働省: 疾病、傷害及び死因の統計分類, <http://www.mhlw.go.jp/toukei/sippe/index.html>
- [5] U.S. National Library of Medicine: Medical Subject Headings, https://www.nlm.nih.gov/mesh/intro_preface.html
- [6] R. Agrawal T. Imieliski, and A. Swami: *Mining association rules between sets of items in large databases*. Acm sigmod record. Vol. 22. No. 2. ACM, pp. 207-216, 1993.
- [7] R. Agrawal, and R. Srikant: *Fast algorithms for mining association rules*. Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215, pp. 487-499, 1994.
- [8] R. Srikant, and R. Agrawal: *Mining generalized association rules*. Research Report RJ 9963, IBM Almaden Research Center, San Jose, California, 1995.
- [9] ヤフー株式会社, 国立情報学研究所 情報学研究データリポ

- ジトリ: Yahoo!知恵袋データ(第3版), https://www.nii.ac.jp/dsc/idr/yahoo/chiebk3/Y_chiebukuro.html
- [10] 新本拓也, 湯本高行, 金子周司, 磯川悌次郎, 松井伸之, and 上浦尚武: 身体部位の表現の違いを考慮したQAサイトからの病訴の検索. 情報処理学会研究報告データベースシステム (DBS) 2018.18, 1-6. 2018.
 - [11] 京都大学大学院情報学研究所黒橋・河原研究室: 日本語形態素解析システム JUMAN, <http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN>
 - [12] 工藤拓: CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer <https://taku910.github.io/cabocho/>
 - [13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov: *Enriching Word Vectors with Subword Information*. Transactions of the Association for Computational Linguistics 5, pp.135-146, 2017