Vol.2018-DBS-167 No.18
Vol.2018-IFAT-132 No.18
2018/9/13

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

新本 拓也[1]　湯本 高行[1]　金子 周司[2]　礒川 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 QA サイトから自身の症状に似た症状についての質問と回答を検索するための手法を提案する．まず，検索対象のデータベースの構築では，QA サイトの質問記事のうち，病訴を含む文を判定し，そこから病訴や部位などの詳細情報を抽出して質問との対応関係を格納する．検索においては，病訴および病訴の現れた部位をクエリとし，入力されたクエリまたはその類義語を含む質問を前述のデータベースから検索する．さらに，部位についてはより詳細な部位の推薦を行うことで，検索意図により合致したクエリへの修正を促す．たとえば，「足 痛み」を検索した場合であれば，「足首」や「踵」などを推薦する．また，類義語については専門用語と日常用語の双方に対応するため，ライフサイエンス辞書と Wikipedia を併用して求める．

1. はじめに

QA サイトの質問記事は膨大で多様な人々が日々書き込みを行っている．そのため，専門家が書くようなテキストに比べて，表記揺れの種類や幅が多くなり，キーワードによる完全一致検索だけでは望む記事を見つけることは難しい．特に医療系の質問記事のテキストでは電子カルテなどの医療分野でのテキストに比べて身体部位に表記揺れが多いため，ユーザが使いやすいように表記揺れを統一するような仕組みや，検索したい情報を絞り込むためのキーワードを提示できるような手法が求められる．

本研究では質問記事のテキストから抽出した病訴を用いて係り受け解析による詳細の抽出を行い，それを元に QA サイトから病訴を検索できる手法を提案する．また，部位の包含関係を表現したツリーを使って，段階的にテキストを絞り込んでいく手法も提案する．提案手法によって，ユーザに対して大まかな部位からより詳しい部位を提示することができたり，ユーザが「ちくちくするするような」といった自分が感じているものに近い状態での病訴のテキストを容易に探し出せるようになる．

2. 関連研究

2.1 MedNLP

本研究の関連研究として医療用言語処理のコンテストである NTCIR-10 の MedNLP[1] がある．その主なタスクとして，電子カルテに年齢や日時，かかった病院などと言った情報を示す各タグを付与するタスクと電子カルテに症状やその診断を示すタグを付与するタスクがある．

後者の症状やその診断についてタグを付与するタスクは本研究で行う病訴の抽出に近いが，テストデータとして配布されるカルテデータは小規模な疑似カルテに留まっているため，本研究の QA サイトを使うものよりデータ数が少ない．また，MedNLP の前者のタスクは日時などといった詳細を抽出する点で，本研究で行う病訴の詳細の抽出と似通った情報を抽出する．しかし，前者のタスクではカルテからの個人データの抽出に重きが置かれているが，本研究で行う病訴の詳細の抽出はテキスト中の対象の病訴の状況や様態により焦点を当てたものとなっている．

2.2 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しての類語辞典で，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

[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

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]．LSD に収録されたこの概念ツリーを利用し，検索を行う．

3. 提案手法

3.1 病訴と詳細の抽出

まず，検索システムを構成するために以下の 3 ステップで病訴とその詳細の抽出を行う．

(1) 病訴を含むテキストの分類
(2) 病訴の抽出
(3) 係り受け解析による詳細の抽出

(1) では個々の質問記事のテキストから病訴を含む文を選別し，(2) ではそのテキストから病訴の抽出を行う．(3) では係り受け解析器を用いて病訴の様態や自覚した状況などを抽出し，病訴についてより正確に把握する．

3.1.1 病訴を含むテキストの分類

QA サイトから得られる質問記事のテキストは病訴に関係ないものが含まれている場合も多い．したがって，このステップでは病訴を含まないテキストを除去する作業を行う．

ICD-10 を使用し，その日本語訳として MEDIS:ICD-10 対応標準病名マスター，LSD を用いた完全一致検索でテキストを病訴を含むものと含まないもので分類する．

分類を行う際には**表 1** に示すパターンに当てはまるテキストは除外する．「よう」,「らい」は病訴以外として頻出する単語であり，「不安」は不安症という病状と混同されるため，この三つの候補はテキスト中に一致する部分があった場合候補から除外する．「うつ」は，病訴としてよく出てくるものの，病訴以外でも「うつる，うつす」としても文章中で一致することが多い．したがって，「うつ」という病訴が出てきた際には，文節の形態素を原形に戻し，それが「うつす，うつる」だった場合には候補から除外する．「心配」は，「心配症」という病訴の場合もあるが，多くの場合は「○○で心配です」などの形で現れる．そのため，「心配」という単語が出てきた文節において「心配症」という単語になっていない時，除外する．また，病訴でない判定として「(候補) でない」「(候補) でしょうか」という文がテキスト内に存在する場合に除外する．さらに，痛み止め，風邪薬などの「(候補) 止め」「(候補) 薬」「(候補) の薬」の場合も，病訴とは異なるため除外する．そのほかの候補に関しても，文節で区切った際に候補が文節の境界にまたがる場合は除外する．たとえば，「う歯」は「そういう歯」という文と一致する．このような場合に，文節で区切ることによって「そういう/歯」となり，「う歯」とは一致しないようになる．

表 1 テキスト分類の除外ルール

病名	条件
よう，らい，不安	テキスト内で一致
ブラ	テキストに「ブラシ」を含む場合
うつ	語を原形に戻した際に「うつる，うつす」
心配	文節内で「心配症」が含まれない場合
「(候補) でない」 「(候補) でしょうか」	テキスト内で一致
(候補) 止め (候補) 薬 (候補) の薬	テキスト内で一致
その他病名	文節で区切った時に一致しない

表 2 病訴抽出時のコスト設定

対象	コスト
名詞 [症]	1
名詞・接頭詞・動詞	6
その他	3

3.1.2 病訴の抽出

病訴の抽出においては処理の高速化のため，テキストから大まかに病訴の候補を絞った後に，候補の中から病訴とその対応する部分を抽出する 2 ステップで行う．

(1) テキストからの病訴候補の決定
(2) 候補からの病訴の抽出

テキストからの病訴候補の決定ではテキストと病訴のそれぞれの名詞の集合から，テキスト内にある可能性の高い病訴の候補をリストアップする．そのために，病訴とテキストをそれぞれ形態素に分解して名詞の集合を作り，共通の名詞数からスコアを計算する．スコアの降順で順位を付け，上位 20 件を病訴の候補と決定する．スコアは以下の式で定義される．

$$score = \frac{\text{病訴と対象のテキストの共通の名詞数}}{\text{病訴の名詞数}}$$

たとえば，「軽い腹痛や腰痛などがあります」というテキストならば名詞は「腹痛，腰痛」の 2 つとなる．ここで「腰痛症」という病状ならば「腰痛，症」で名詞数は 2 つである．テキストと病訴の共通の名詞数は「腰痛」だけなので，スコアは 1/2 で 0.5 となる．このスコアが大きいほど順位は高くなる．

(2) ではこの候補からテキスト内の病訴に該当する部分を見つけて抽出を行う．形態素を単位としたウィンドウの大きさを変えていき，抽出した部分と (1) で決定した病訴候補を比較する．病訴を抽出できた周辺でウィンドウサイズを 1 から大きくして 8 まで変化させ，ウインドウごとに病訴の候補と一致しない形態素のコストの合計を計算する．. 合計のコストが最小となったウィンドウ位置を関連性の高い病訴部分と判定する．

コストは**表 2** のように設定する．たとえば，「軽い腹痛

や腰痛などがあります」という文でウィンドウ幅を 1，2 とすると，それぞれ「腰痛」と「腰痛/など」といった部分が取れる．それぞれに対して「腰痛症」といった病状とのコストを考えると，前者では共通しない形態素は「症」だけとなりコストは 1 となる．後者では「症/など」が共通しない形態素となり，「症」はコスト 1，「など」は助詞でコスト 3 なので合計のコストは 4 となる．この場合ウィンドウ幅が 1 の場合の「腰痛」の部分が，病状と最も関連性の高い病訴部分だと抽出される．

3.1.3 係り受け解析による詳細の抽出

抽出した病訴に関連する詳細の情報を文章から探して抽出を行うことで病訴のより正確な状態を把握して提示する．対象とする詳細のカテゴリは「状況」，「部位」，「様態」の 3 つとする．「状況」は時間と場所で判定し，時間は起こった場所や病訴の期間，場所は起こった地点や病訴が判定された場所を示す．「部位」は病訴が現れている身体の部位を示す．「様態」は病訴がどのようになっているかの状態を示す．たとえば「痛み」に対して「チクチクとした」といった具体的な感じ方などが該当する．

詳細の抽出はまず，病訴を抽出した部分から係り受け解析によって詳細を抽出する文節を決定する．係り受け解析には JUMAN[6] で形態素解析した結果に対して KNP[7] で係り受け解析をする．詳細として抽出するために使用する文節は以下のように辿っていく．

(1) 病訴の含まれる文節の係り元の文節
(2) 病訴の含まれる文節の係り先の文節
(3) (2) の係り元の文節やその係り元を辿っていった文節

例として「5 日前から 37 ℃台の発熱が出現した」というテキストの詳細を抽出する場合に使用する文節を**図 1** に示す．また，係り元を辿っていく際に文節の末尾が動詞で終わっていたらそれ以上は辿らない．これによってたとえば「5 日前から発熱が出現し，3 日前から腰痛が出現した」というテキストがあった場合に，「腰痛」という病訴と「5 日前」という誤った詳細を結び付けることがなくなる．

続いて「状況」，「部位」，「様態」の 3 つのカテゴリについて，文節に**表 3** に示すパターンのいずれかの情報が含まれていた場合にそれを各詳細と判定する．「状況」の時間では，KNP で解析した際に付与される分類において「時相名詞」もしくは「カテゴリ：時間」である時，または「歳」「時々」「(名詞)+中」のいずれかがが文節中に含まれているときに判定する．「(名詞)+中」はたとえば「休憩中」などのことである．「状況」の場所では KNP で解析した際に付与される情報に「カテゴリ：場所」がある場合に判定する．「部位」では KNP で解析した際に付与される情報に「カテゴリ：動物-部位」が含まれる場合に判定する．加えて，LSD の概念ツリーから部位の情報を抽出したものを使用して，それに含まれるものが文節中に存在する場合に判定する．「様態」では文節内に「〜のような」といった文

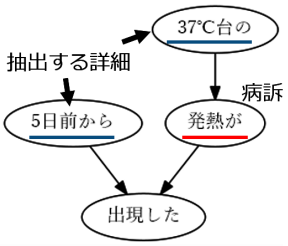

図 1 詳細抽出

表 3 詳細判定の条件

詳細		判定条件
状況	時間	・時相名詞 ・「歳」「時々」「(名詞)+中」のいずれか ・カテゴリ：時間である
	場所	・カテゴリ：場所である
部位		・カテゴリ：動物-部位である ・LSD 概念ツリーの部位情報のいずれか
様態		・「〜ような」 ・形容詞，オノマトペ，カテゴリ：色 ・病名が「高熱」時に「℃/度」

むずむず, がさがさ, いらいら, ごろごろ, ぶつぶつ, ぴりっ,
かさかさ, ぜーぜー, いがいが, ちくりちくり, さらさら, どろどろ,
ぼんやり, ぶよぶよ, ずっと, ぼうっ, ぼーっ, どろっ, ちくっ,
ぐるぐる, ぐるりぐるり, ぐったり, ぴりぴり

図 2 Yahoo 知恵袋から得たオノマトペ

章，形容詞やオノマトペが含まれる場合，KNP で解析した際に付与される情報に「カテゴリ：色」を含む場合に判定する．この時にオノマトペはメディカルオノマトペ [8] から 22 語と**図 2** に示す Yahoo!知恵袋から抽出した病訴に関わりのありそうなオノマトペ 23 語を加えて使用している．特別な場合として，病訴が「高熱」の時に「℃，度」を含む文節があるときにも判定を行う．

また，**表 4** に示すパターンに当てはまる場合は詳細から除外する．「前半/後半」，「右目/左目」は予備実験で誤判定が多かったため，判定の条件から除外する．「度」は温度の場合も，KNP では「たび」と解析され時間の条件になるため，「状況条件」の判定の条件から除外している．また，部位の場合は部分一致検索だと「手術」「牛乳」などもそれぞれ「手」「乳」といった部位の条件に当てはまる．したがって，部位の判定の際，文節を形態素分解して形態素と判定に用いる部位情報と一致しない場合に判定から除外する．たとえば，「手に怪我をした」，「牛乳で腹痛になった」といった文章では「手に」「牛乳で」の文節が条件に当てはまるが，それぞれを形態素に分解した場合，「手/に」「牛乳/で」となるので，「手」は部位の詳細と判定し，「牛乳」は判定から除外することができる．

3.2 質問記事の検索

QA サイトの質問記事テキストを検索しようとすると部

表 4　詳細抽出の除外条件

詳細	除外条件
状況	・「前半/後半」「右目/左目」「度 (温度)」のいずれかを含む文節 ・部位と判定された文節
部位	・形態素分解したときに形態素と判定に用いる部位情報が完全一致しない
状況 部位 様態	・文節に他の病状が含まれる文節

位や病訴などの表記揺れが多く，キーワードの完全一致検索では難しい．したがって部位などを上位下位関係などで体系的に整理してテキスト情報に結びつけたり，次に検索するキーワードを提示できるようなシステムを作ることで，ユーザが検索システムを利用するときに検索したい質問記事を絞り込むことができる．

以降では，部位の上位下位関係を整理する手法と，それを利用したクエリの処理について述べる．

3.2.1　質問記事 DB と部位の概念ツリー

検索のため，質問記事ごとに 3.1 で抽出した情報を格納したデータベース (質問記事 DB) を構築する．質問記事 DB には ID，テキスト，抽出した病訴，ICD コード，状況情報 (時間と場所)，部位情報，様態情報を格納する．また，部位については LSD の概念ツリーの部位に関する部分木を基にした概念ツリーと対応づけることで，類義語や上位下位概念による検索も可能にする．

詳細抽出で現れた部位情報と LSD の概念ツリーとの対応付けは以下のように行う．

(1) LSD の類語関係を使用した完全一致検索
(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検索
(3) 既に関連づけられた部位情報との部分一致検索

(1) は LSD に含まれている類語関係を使用して，概念ツリーのそれぞれのノードに質問記事をデータベースの部位情報から対応させる段階である．LSD の概念ツリーは語の包含関係をツリーとして表示したものであり，たとえば「下肢」という語の下位に「足」や「膝」といった語が配置される．

(2) は各ノードに結びつけられている部位から Wikipedia のリダイレクト・リンクアンカー関係を調査した際に現れた単語をその部位に関連する単語とみなして対応づける．その主な対象は日常的な用語である．Wikipedia から得られるリダイレクトは記事名の略称や別表記を実際の記事に転送する機能であり，リンクアンカー関係は記事の本文中の語からその語の記事にリンクを付ける機能である．ある単語を入力した際に Wikipedia に登録されている記事ならばリダイレクトで転送されてくる単語がリダイレクト元として現れ，リダイレクト元の単語ならば転送先の単語がリダイレクト先として現れる．このリダイレクト元，あるいはリダイレクト先に概念ツリーに対応されている単語が現れたなら，概念ツリーに対応づける．また，ある単語が Wikipedia 内の記事のアンカーテキストとしてある記事にリンクされている場合のリンク先の記事名と，ある単語が Wikipedia 内の記事としてリンクされている場合にそのアンカーテキストとなる単語も同意語とみなして概念ツリーに対応づける．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．ここにリダイレクト関係を適用すると「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．

(3) はたとえば「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ける．

3.2.2　クエリの処理

QA テキストの検索のクエリとしては病訴，部位を想定する．複数のクエリで検索を行う場合，それぞれのクエリの検索結果の質問記事の積集合を取る．

病訴がクエリとして入力された際の処理は以下の 3 ステップで行う

(1) 3.2.1 のデータベースの病訴から部分一致検索で質問記事のテキストを発見する．
(2) 発見したテキストから対応する ICD コードを使い，ICD コードの共通する他の病訴の質問記事のテキストも検索結果に含める．
(3) 検索結果となるテキストのそれぞれの病訴をクエリに対する類病として，それぞれのテキストと対応する様態情報を関連する状態としてユーザに提示する．

部位がクエリとして入力された際の処理は以下の 3 ステップで行う．

(1) 入力された部位に対して，3.2.1 の概念ツリーから対応する部位のノードを見つける．
(2) 発見したノードに対応した部位が部位情報に含まれるテキストを 3.2.1 のデータベースから抽出し，検索結果とする．
(3) 発見したノードの子ノードから得られる部位を関連部位として表示する．

この検索システムによって，検索結果ごとに類病と関連する状態の提示ができ，病訴と様態ごとの検索結果の絞り込みができる．また，部位情報を大まかな部位から細かい部位への絞り込みも行うことができる．

表 5　病訴を含むテキストの分類の実験結果

適合率	再現率	F 値
500 文/738 文 = 0.678	500 文/595 文 = 0.840	0.750

4. 実験

4.1 病訴，詳細の抽出

4.1.1 病訴を含むテキストの分類

3.1.1 の手法により Yahoo!知恵袋の医療に関連する記事として健康・美容とファッションの健康カテゴリの質問記事からランダムに選んだ 500 件に含まれるテキスト 2827 文を分類した．その結果から得られた適合率，再現率，F 値を**表 5** に示す．

表 5 を見ると再現率は比較的高いが，適合率は低くなっている．これは分類の際の条件だけでは病状に対して疑問や否定を示している文を除ききれなかったためと考えられる．

4.1.2 病状の抽出

3.1.2 の手法により，病訴を含むと分類された質問記事のテキスト 306 文について病状を抽出し，Yahoo!クラウドソーシングを用いて正解を判定した．クラウドソーシングでは，病状を抽出したテキストと抽出された病状を提示して，病状がテキスト中に含まれているか否かを分類してもらった．1 文あたり 5 人に分類を行ってもらい，多数決で正解を判定した．その結果適合率は 254 文/306 文 = 0.830 となり，8 割程度の適合率で抽出ができていることが分かった．

不正解だったものでは「手術」「切除」といった ICD-10 対応標準病名マスターや LSD に登録されているが病状でないもの，「脂肪」→「脂肪症」といった一部の単語だけに反応して病状と判定してしまうものがあった．

4.1.3 係り受け解析による詳細の抽出

3.1.3 の手法により質問記事のテキストから抽出した病状で，詳細の付随するものの詳細の適合率を調査した．調査手法として抽出した症状の「状態」，病状の起こっている「部位」，いつどこにいる時病状を自覚したかの「状況」のそれぞれに対してクラウドソーシングを用いた．クラウドソーシングの参加者に質問記事のテキストとそのテキストから抽出した病状，その病状の詳細それぞれを提示して，その詳細が病状に対して正しいかどうかを分類してもらった．「状態」の詳細が含まれるテキストから 116 文，「部位」が含まれるテキストから 120 文，「状況」が含まれるテキストから 120 文を 1 文あたり 5 人に分類を行ってもらい，多数決で正解を判定した．その結果を**表 6** に示す．

状態，部位，状況のそれぞれの結果を見ると，それぞれ 7 割以上の適合率を実現できた．不正解のものを見ると，係り受け解析の際に「ふわぁとしたような」や「胸あたりの背骨」，「若い頃」など複数文節を考慮していないため，抽出した詳細がそれ単体では意味の分からないものになってる場合もあった．したがって，複数文節を考慮した抽出を行うことで，より正確な詳細情報の提示を行うことができると考えている．

表 6　詳細の抽出の実験結果

	正解数	適合率
状態	84 文/116 文	0.724
部位	93 文/120 文	0.775
様態	98 文/120 文	0.817

表 7　概念ツリーとの対応付け

手法	件数
LSD の類語関係を使用した完全一致検索	138 語
+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検索	205 語
+既に関連づけられた部位情報との部分一致検索	228 語
概念ツリーに関連づけられなかったもの	13 語

4.2 質問記事の検索

4.2.1 部位の対応付け

抽出した部位情報を LSD の概念ツリーに対応づける際に，3.2.1 の三段階の関係性を用いて対応づけた．たとえば 6079 文の質問記事のテキストから「痛み」という病状に関して詳細の抽出を行った結果，241 語の部位情報が得られたが，この 241 語に対して三段階で概念ツリーに対応付けた結果を**表 7** に示す．

LSD の類語関係と Wikiepdia のリダイレクト・リンクアンカー関係を組み合わせることで，241 語のうち約 95%を LSD の概念ツリーに対応づけることができた．概念ツリーに対応づけられなかったもののうち，「おやしらず」のような今回の手法では分類できなかったり，「全身」といった概念ツリー上に分類する箇所がなかったものは 5 件あり，部位情報が含まれていない抽出段階での不正解と考えられるものは 8 件あった．

4.2.2 クエリの実行例

この検索システムで「痛み」という病訴を検索したとき，1612 件のテキストが出力される．この検索によって「痛み」の類病として ICD コードを共通する病訴として「鈍痛」，「疼痛」，「痛」なども出力される．また，状態として 247 件が出力され，その中には「ズキッと」，「鋭い」，「強い」，「微妙な」などが含まれる．状態の語彙に関してはまだ統合されておらず，「ズキズキ」や「ズキッ」といった語彙を一つにまとめる手法を開発することが今後の課題となる．

部位に関する例として「手」というクエリを検索した場合，173 件の検索結果が得られる．この「手」というクエリには概念ツリーで手というノードに結び付けられた「右手，左手，両手，手のひら」などのクエリの検索結果も含

	text	symp
0	ちなみに私は、左下腹部に何か詰まっているような痛みや、同じく左下腹部がチクチクと痛いときがあります	痛み
1	最近左下腹部がときどき痛いです	痛
2	ポリープを切除したのに、まだ、おなかがちくちく痛いです	痛
...	...	...
90	いつも僕は学校行くと急に腹が痛くなります何も薬を使わなくて急な腹痛を治す方法とかありますか	痛
91	おへそから下あたりが痛く手で押さえたらさらに痛いのですが考えられる病気は何があるでしょう	痛
92	昨日から左のへそ回りが痛いのです	痛

93 rows × 2 columns

類病 ['痛', '痛み']
関連部位: ['腹腔', '網', '臍']
状態 ['詰まっているような', '絞られるような', '多くなってから', '激しく', 'キリキリと', 'ずっと', 'チクチク、', 'チクチクと', '強い', '急に', '普通に', '軽い', '真ん中辺りで、', '高速', 'キリキリ', 'ぐるぐる', '吊られるような', '大量の', 'するような', '強く', '締め付けられるような', '悪くなったのと、', 'あるような', 'すっごく', 'ズキズキと', '一気', '突然、', 'ないんですが、', 'きれいな', '重く', 'ムズムズ']

図 3 「腹，痛み」の検索結果

まれる．「手」という検索結果ではさらに関連部位として「手指，おや指，手首」といった下位概念も出力されるため，さらに絞り込みが行える．

ここで，例として「痛み，腹」と検索した場合の検索結果を図 **3** に示す．検索結果として 93 件のテキストが出力され，類病として「痛，痛み」，関連部位として「腹腔，網，臍」が出力されている．類病は「痛み」と検索した際に，同じ ICD コードで結び付けられた病訴が出力されている．関連部位には 3 つが出力されているが，このうち「網」は omentum，つまり大網や小網のことであり，「腹腔」の下位概念でもある．関連部位が複数出てくるようなクエリ検索では下位概念の検索結果も含むため，出力される検索結果が多くなる．ここで関連部位の単語をクエリとして続けて入力することで，より細かく検索結果を絞り込める．

5. おわりに

本研究では，QA サイトの質問記事のテキストから病訴とその詳細を状況，部位，状態の三要素で抽出する手法を提案した．また，QA サイトの質問記事の検索法として抽出した詳細を利用する方法を提案した．

病訴の抽出においてはテキストに病訴が含まれるかどうかを分類してから形態素ごとにコストを設定して病訴を抽出し，詳細の抽出においては係り受け解析を用いてルールベースの抽出を行った．質問記事の検索システムの構築では，LSD の概念ツリーを用いて部位同士の関係を関連づけ，検索の絞り込みが行いやすいような手法を提案した．

評価実験においては病訴の抽出では 8 割程度の適合率で病訴を抽出でき，詳細抽出においても 7 割以上の適合率で抽出できたが，複数文節を用いた抽出など改善の必要がある点も確認できた．また，検索システムを構築時の部位の概念ツリー作成においては抽出できた部位情報のうち，大部分の語彙を LSD の概念ツリーと関連づけることができた．今後の課題として，抽出手法の改善と，検索システム構築時に部位だけでなく状態なども統合できる手法の開発が挙げられる．

謝辞 本研究の一部は平成 30 年度科研費基盤研究 (C)(17K00429) によるものである．

参考文献

[1] Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, Mai Miyabe, Eiji Aramaki: *Overview of the NTCIR-10 MedNLP task*, In Proceedings of NTCIR-10, pp147-154, 2013.
[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/sippei/index.html
[5] U.S. National Liberary of Medicine: Medical Subject Headings，https://www.nlm.nih.gov/mesh/intro_preface.html
[6] 京都大学大学院情報学研究科黒橋・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
[7] 京都大学大学院情報学研究科黒橋・河原研究室: 日本語構文・格・照応解析システム KNP，http://nlp.ist.i.kyoto-u.ac.jp/?KNP
[8] 国立国語研究所，株式会社オズマピーアール，小野正弘: オノマトペラボ・メディカルオノマトペ，http://onomatopelabo.jp/medical/gram/index.html