一般口演 | 知識工学
# 一般口演10
# 知識工学
2019年11月23日(土) 09:00 〜 11:00 C会場 (国際会議場 2階国際会議室)

[3-C-1-05] 句構造を取る症状表現を大規模 Webテキストから取得する試み

○和田 聖哉[1]、飯田 龍[2]、鳥澤 健太郎[2]、武田 理宏[1]、真鍋 史朗[1]、小西 正三[1]、松村 泰志[1] （1. 大阪大学大学院医学系研究科医学専攻 情報統合医学講座医療情報学, 2. 情報通信研究機構）

キーワード：Natural language processing, Machine learning, Symptom extraction

自動診断を始めとした診断支援システム開発には疾患名と症状が対応付けられた大規模な辞書が必要である。「咳嗽」「頭痛」といった単語レベルの症状辞書は存在するが、「膝が痛い」「耳鳴りが続く」といった句レベルの症状については、クラウドソーシングを用いて構築された患者表現辞書（ MedNLP）内に散見される程度である。患者の自然な発語から疾患名を推定するには、それら句レベルの症状辞書が必須である。本研究では、大規模 Webテキストに自然言語処理技術を用い、句レベルの症状表現を効率よく取得する手法を提案する。

本研究では、「頭が痛い」のような、＜名詞＋助詞＋述語＞からなる表現を対象とする。＜Aで Bが起こる＞のように、述語とそれに係る項（ Aと Bは名詞句）が2つの言語パタンをバイナリパタンと呼ぶ（以降、 P）。大規模 Webテキストに構文解析を行い、言語パタン Pと付随して出現した A, Bの共起頻度(P, A, B)を求めたリストを準備した（約180億パタン）。 (P, A, B)のリストから Aもしくは Bに疾患名が入り、その後の助詞が「で」になるものを抽出した（疾患名は ICD10対応標準病名マスター等から取得した約2万語を使用。全426,975パタン、最頻出は「発疹がひどい」で653,647）。取得した表現の名詞について、①疾患名もしくは症状、②身体部位に分け、それぞれに共起した「助詞＋述語」を症状テンプレートとした。症状テンプレートを元に、①以前我々の提案した症状表現抽出手法で取得したリスト、② ICD10対応標準病名マスター修飾語テーブルを単語集合作成ツールに適用して取得した身体部位リストでそれぞれ拡張した。それらを(P, A, B)のリストと照合し、 Web上で使われた表現かどうかを確認した。

拡張した表現は Webテキスト上での使用頻度を確認した段階に留まる。実際に症状表現として妥当かどうか、人手評価を検討している。

句構造を取る症状表現を大規模 Web テキストから取得する試み

和田 聖哉[*1]、飯田 龍[*2]、鳥澤 健太郎[*2]、武田 理宏[*1]、真鍋 史朗[*1]、小西 正三[*1]、松村 泰志[*1]
*1 大阪大学大学院医学系研究科 医療情報学、*2 情報通信研究機構

Approach to acquisition of phrasal symptom expressions from a large web archive: A pilot study.

Shoya Wada[*1], Ryu Iida[*2], Kentaro Torisawa[*2], Toshihiro Takeda[*1],
Shiro Manabe[*1], Shozo Konishi[*1], Yasushi Matsumura[*1]
*1 Department of Medical Informatics, Osaka University Graduate School of Medicine,
*2 National Institute of Information and Communications Technology

We propose a method to get phrasal symptom expressions using a large web archive that includes a large amount of texts written by non-medical experts. Our final goal is to develop a diagnosis support system that makes a diagnosis according to the natural language inputs provided by patients. Existing work for symptom expression extraction have mainly focused on the extraction of "symptom words", but in this work we focus on the extraction of "phrasal symptoms", such as “have a pain in the head” and “have tinnitus”, which were less focused on in the previous work. In the proposed method, we extract phrasal symptom candidates that contain a word representing a body part or a symptom word that was collected in our previous work. To do that, we developed a Bidirectional Encoder Representations from Transformers (BERT)-based body-part classifier using manually created training instances. We assessed the usefulness of our phrasal symptom extraction method by ranking the extracted phrasal symptom expressions with their frequencies and manually evaluating the top 2,000 expressions. The results show that our method achieved 78.9% in accuracy.

Keywords: Natural language processing, Machine learning, Symptom extraction

1. 緒論

自動診断をはじめとした診断支援システム開発には疾患名と症状が対応付けられた大規模な辞書が必要である。日本語のリソースでは、「咳嗽」「頭痛」といった名詞単語レベルの症状辞書 [1, 2]は存在するが、「膝が痛い」「耳鳴りが続く」といった句レベルの症状表現（以後、「句症状表現」と呼ぶ）については、MEDNLP が公開している患者表現辞書 [3]に数千件の規模で含まれるが、患者との自由なやりとりの中に出現する多様な症状を十分に網羅できているとはいえない。

患者表現辞書の情報源（エビデンス）は、クラウドソーシング（CS と表記、6,148 レコード）、患者症状抽出器 Adverse Effect Extractor[4]（AEX：ルールベースによる症状表現抽出ツール、2,144 レコード）、Web（全てオノマトペを含む表現で構成、570レコード）とされており、機械学習を利用した句症状表現収集手法に関しては、我々の知る限り報告がない。

そこで、本研究では過去に提案した大規模 Web テキスト集合から症状単語（以後、「症状名詞」と呼ぶ）を獲得する手法を拡張し、句症状表現を獲得する手法を提案する。

2. 目的

患者が疾患に罹患した場合、その症状を表すのに用いられる句症状表現の多くは、「頭が痛い」「頭痛がする」「熱が出る」「鼻が詰まる」のように、名詞と述部から成る。1 つもしくは複数の単語から成る症状単語(例：「咳」「頭痛」)と比較して表現の総数が多くなることが予測されるため、本研究ではまず句症状表現の名詞部分に身体部位を表す名詞もしくは症状名詞を含む場合を対象に句症状表現を抽出し、その抽出の性能を評価し、さらに表現の多様性を担保するために今後どのような取り組みを行うべきかを検討する。

3. 方法

図 1 に提案する句症状表現獲得手法の概要を示す。本手法の核となる技術は大規模 Web 文書からの効果的な句症状表現の抽出であり、このために我々は大規模情報分析システム WISDOM X (https://wisdom-nict.jp/)と呼ばれる約 40 億件の Web 文書を対象とした質問応答エンジンを利用する [5]。我々がこれまでに行った症状名詞の獲得でも WISDOM X を利用した症状名詞の候補の抽出を行っており [6]、句症状表現の候補抽出でも類似する技術を用いることを試みる。

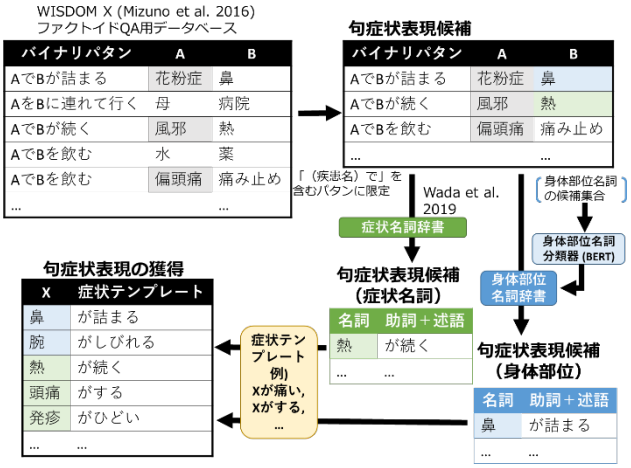

図 1 本研究の概要

具体的には、WISDOM X のファクトイド質問応答(QA)システムで利用されるファクトイド QA データベースから句症状表現候補を獲得し、次に句症状表現の名詞部分（例えば、「膝が痛い」の「膝」の部分や、「吐き気がする」の「吐き気」の部分）に着目し、名詞部分が「膝」のような身体部位である場合、もしくは、「吐き気」のような症状表現である場合に限定して句症状表現を獲得することで、最終的な句症状表現を獲得する。

以降でまず句症状表現の候補獲得について説明し、次に提案する身体部位を表す名詞(以降、「身体部位名詞」と呼ぶ)の獲得手法とその評価結果を報告する。さらに、獲得した身体部位名詞もしくは症状名詞を含む句症状表現候補にど

の程度適切な句症状表現が含まれているかの評価を、比較対象となる患者表現辞書と比較を行いながら評価を行う。

3.1 句症状表現候補の収集

＜A で B が起こる＞のように、述語とそれに係る項（A と B は名詞句）が 2 つの言語パタンをバイナリパタンと呼ぶ（以後、P と表記する）。また、簡単のために、「が起こる」に該当する「助詞＋述語（動詞、形容詞、形容動詞などの用言）」を症状テンプレートと呼ぶ。WISDOM X[5]のファクトイド QA 用データベースに登録されている、言語パタン P と付随して出現した A, B の共起頻度(P, A, B)を求めたリスト（約 180 億パタン）から、句症状表現の候補を抽出する。Web 上で患者が句症状表現を記述する際には、その原因（疾患名）とともに記入する可能性が高いと考えられる（例：「風邪で熱が出る」、「蕁麻疹で皮膚が痒い」）。このような原因を表す助詞は同データベース内では「で」に標準化されているため、 (P, A, B)のリストから A もしくは B に疾患名が入り、それに続く助詞が「で」になるものを条件にして抽出した。疾患名には、平成 30 年版医師国家試験出題基準 [7]、ICD10 対応標準病名マスター V 4.02[8] 、ALAGIN 文脈類似語データベース [9]から取得した単語で構成される疾患名リスト（19,273 語）を使用した。これにより、426,975 パタン（名詞数 196,651、症状テンプレート 4,621）を取得した。このうち、共起頻度が 5 以上であるものを本研究の対象とした（89,158 パタン、名詞数 41,717、症状テンプレート 3,315）。取得結果の一部を表 1 に示す。本研究で収集対象にしているのは句症状表現であるが、このような表現とともに、表 1 で示すような、句症状表現として「好ましくない事例」が少なくない頻度で対象の中に混在している。

表1　句症状表現とそれ以外の表現の例

句症状表現	発疹がひどい，熱を出す，喉をやられる，体調を崩す，熱が出る，頭痛がある，歯が溶ける，意識を失う，腰が曲がる，喉が痛い，記憶を失う,…
好ましくない事例	病院に行く，手術を実施する，妻が倒れる，辞退者が出る，食事がする，脂肪燃焼が低下する，入退院を繰り返す，療養を受けられる，薬を飲む，2月に倒れる,…

3.1 で収集した 89,158 パタンに含まれる名詞から 400 事例をサンプリングして調査したところ、表 2 の結果となった。

表 2　対象パタンに含まれる名詞の分類とその割合

症状	7.25%
身体部位	2.00%
その他	90.75%

表 2 の中で、その他の名詞を用いる形では句症状表現は認められなかった。このサンプリング結果を受けて、本研究では 2 節で述べたように、1. 身体部位（例.「頭」が痛い、「腰」が曲がる、など）、2. 症状（例.「高熱」が出る、「下血」が続く、など）が名詞部分に含まれる表現を句症状表現の候補とした。

このような表現から目的とする句症状表現を抽出するために、対象とした表現中に存在する名詞の分類から句症状表現を獲得する手法を提案する。本提案手法は以下に従う：

1. 句症状表現候補内の名詞を利用したフィルタリング（分類モデルの構築と名詞辞書とのマッチング）
2. 名詞の出現頻度・症状テンプレートの出現頻度による対象表現のランク付け

3.2 句症状表現の抽出

身体部位、症状名詞それぞれの評価・取得方法と、それによる句症状表現の抽出手法について説明する。

3.2.1 身体部位名詞の獲得とその評価

2 節で述べたように、身体部位名詞を網羅的に含む辞書が存在しないため、それをまず作成する必要がある。ここでは辞書に含まれうる候補を網羅的に人手で記述するのではなく、少量の身体部位名詞の候補にアノテーションを行い、そのアノテーション結果を用い、機械学習による分類器を作成し、その分類器を大規模な身体部位名詞候補集合に適用することで大規模な身体部位名詞の辞書を作成する。この分類器には近年さまざまな自然言語処理のタスクにおいて高い性能を得ている Bidirectional Encoder Representations from Transformers (BERT) [10]を利用する。

身体部位名詞は「頭」のような一般的な名詞から「耳介筋」のような専門用語までさまざまであるが、自動診療で患者が一般的に述べる身体部位を収集するためには一般的な身体部位名詞を含む情報源から知識獲得を行う必要がある。そこで、学習に用いる身体部位名詞の初期シードには一般的な身体部位名詞も含むと考えられる Wikipedia 日本語版のダンプデータ（jawiki-20190601）を利用した。具体的には、句症状表現の名詞部分に使われる可能性の高い「人体部位・骨・関節・筋肉」を正例とするために、Wikipedia 日本語版のページ「人体解剖学」・「人間の筋肉の一覧」・「人間の関節一覧」・「人間の骨の一覧」とカテゴリ「人体の部位」にある単語を取得した（454 語）。一方で、ミクロレベルの表現はこれらの身体部位の類似語として判定される可能性があるものの、それらが句症状表現に使用される可能性は低いため、同ページの「ヒトの細胞の一覧」とカテゴリ「ヒト細胞」・「組織学」にある単語を初期シードの負例として設定した（312 語）。それらの一部を表 3 に示す。

表 3　身体部位名詞の具体例

正例 From Wikipedia （人体解剖学，人間の筋肉の一覧，人間の関節一覧，人間の骨の一覧，人体の部位）	口，鼻，胃，心臓，腰，胸，肘，上腕二頭筋，膝関節，肩甲骨，…
正例 文脈類似語データベース	背中，お腹，手足，体，あご，ひざ，スネ，すい臓，腸，肌，…
負例 From Wikipedia （ヒトの細胞の一覧とカテゴリ，ヒト細胞，組織学）	赤血球，白血球，破骨細胞，粘膜筋板，ぶどう膜，基底膜，骨膜，柔組織，杯細胞，ニューロン,…
負例 文脈類似語データベース	自分，細胞，窓，装具，プライベートスペース，中、味蕾，コラーゲン，…
負例 Negative sampling From 名詞意味クラス辞書	フロントタイヤ、バンパー、専制国家、位置、中心、山薬、進み具合、煙突、電気石、甲羅、…

次に正例の学習データを増やすために、既存のデータベ

ースを用いて拡張を行う。身体部位初期シードの 454 語を ALAGIN 文脈類似語データベース [9]に適用し、10,271 語の身体部位名詞候補を獲得した。この中から 2,447 事例をサンプリングし、人体部位（外表・内臓）・骨・関節・筋肉か否かで判定を実施した。初期シードの 766 語を含め、合計 3,213 語のラベル付きデータを得た（表 3）。

ここまでに作成した身体部位名詞の候補は、正例・負例のどちらも人体に関連のある単語であると言える。身体部位名詞分類器に入力される身体部位以外の名詞を適切に負例に分類するために、身体部位名詞以外の名詞を対象にした negative sampling を行い学習データに追加する。具体的には、風間ら [10]の単語意味分類手法を大規模 Web 文書（約 6 億文書）に適用して作成された名詞意味クラス辞書を事前に分析し、前出のラベル付きデータに対して適用した上でそれら以外のクラスに属する 20,035 単語を負例に追加した（表 3）。これにより、全 23,248 （正例の割合は 5.8%）からなるデータセットを構築した。これを表 4 のように分割し、学習及び評価に用いた。

表 4　学習・評価用データセットの作成

	負例	正例
train	20,394	1,154
development	1,513	187
Total	21,907	1,341

train データで学習を行い、最適なハイパーパラメータとモデルの選択には、development データを用いた。モデル選択時には、再現率と精度から計算される調和平均 (F1-score)が最大となるパラメータを選択した。

BERT はまず大規模な生テキストで言語モデルなどの目的関数のもとにモデルを事前学習する。そして各タスクで fine-tuning することにより、少ないデータでも様々なタスクで 高い性能を得ている。生化学・医学領域でも、各ドメインで事前学習を行い、タスクの精度が改善した報告がある [12, 13]。

本研究では、日本語 Wikipedia （jawiki-20180801）全文で事前学習した BERT モデル（12-layer、768-hidden、12-heads、dropout 0.1、max_seq_length 768、語彙数 100,004: Juman 辞書を用いた MeCab[14]による形態素解析後）を利用し、fine-tuning を実施した。ハイパーパラメータで学習率（1e-5、2e-5、3e-5、5e-5）、epoch 数（1.0, 2.0, 3.0）を探索し、F1-score 85.1% (Recall 78.3%/ Precision 93.1%)の身体部位名詞分類器を得た。句症状表現候補に用いられている名詞にこの分類器を適用し、学習データに含まれていなかった身体部位名詞の具体例を以下に示す。

右腎臓、満腹中枢、リンパ、気合、股関節、片足、お肌、リンパ節、右足、ピアス、半身、顔色、頭皮、歯肉、上唇、下腹、頭髪、両手首、お鼻、…

本予測モデルを 3.1 節で得た句症状表現候補に含まれる名詞 41,717 語に適用し、その結果 3,506 語が身体部位と判定された。その身体部位名詞を含む句症状表現候補を抽出し、最終的に 19,163 件の句症状表現を獲得した。これらの句症状表現に WISDOM X のファクトイド QA データベースに記録される頻度情報を割り当て、頻度が多い順にランキングした。このランキングの上位群と下位群の比較をするために、1～100 位と 2,001～2,100 位を抽出し、人手で分析したところ、適切な句症状表現の割合はそれぞれ 67.0%、30.0%であった。この結果から、頻度が高い場合に句症状表現となっているものが多いことがわかった。次に句症状表現がより高い割合で獲得できるように、名詞部分と症状テンプレート部分で頻度の集計を行い、その合計数がどちらも 100 以上である句症状表現候補のみを抽出して大規模に句症状表現を獲得することを試みた（以後、「身体部位-句症状表現（候補）」と呼ぶ）。一例を図 2 に示す。これにより、6,657 パタンの身体部位-句症状表現候補を選択した（表 5-3）。

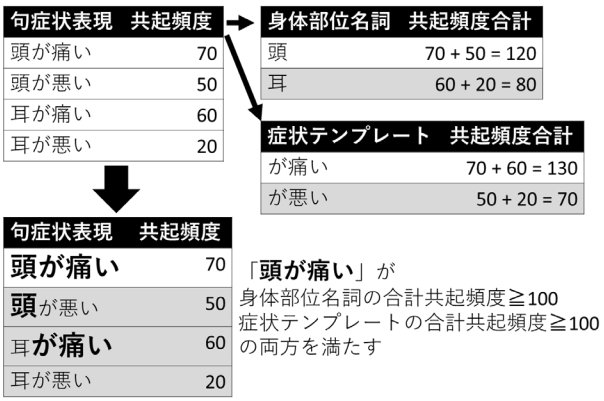

図 2　合計共起頻度による対象句表現候補の抽出手法

3.2.2 症状名詞の獲得

効率的に症状名詞を収集するために、以前我々が提案した、Web テキストから症状名詞を抽出する手法 [6]を適用し、症状名詞リストを作成する。具体的には、WISDOM X[5]に、「（疾患名）で何が生じる？」（例:「風邪で何が起こる」）といった質問の回答を得ることで症状名詞の候補を得る。疾患名約 2 万語を入力して得られた症状名詞の候補の総数は約 350 万件であった。これらに名詞意味クラス辞書 [11]を適用して候補を削減した上で、症状表現抽出モデルを適用し、症状名詞 45,844 語を獲得した（200 事例のサンプリングで、Recall 87.3%/ Precision 89.9%/ F1-score 88.6%の精度）。症状名詞の具体例を以下に示す。

発疹、病気、熱、症状、体調、頭痛、言葉、名前、障害、記憶、高熱、後遺症、咳、可能性、入退院、意識、性欲、体重、出血、短期記憶障害、…

表 5　句症状表現候補の選択プロセス

	処理内容	句症状表現候補数 (名詞数 | 症状テンプレート数)	
		身体部位名詞	症状名詞
1	句症状表現候補の共起頻度≧5	89,158 (41,717 | 3,315)	
2	分類モデルの判定 or 名詞リストとのマッチング	19,163 (3,506 | 2,157)	11,591 (2,992 | 1,390)
3	名詞部分の合計共起頻度≧100 and 症状テンプレートの合計共起頻度≧100	6,657 (195 | 297)	6,353 (691 | 444)
4	句症状表現候補の共起頻度（上位 1,000 まで）	1,000 (191 | 287)	1,000 (494 | 311)

身体部位名詞の場合と同様に、3.1 で得た句症状表現候補のうち上述の 45,844 語の症状名詞を含むものに限定して抽出した。その結果、2,992 語の症状名詞が特定され、その名詞を含む句症状表現 11,591 件が抽出された。名詞部分、症状テンプレートの部分それぞれで合計共起頻度を評価して、ともに共起頻度が 100 以上である句症状表現候補を抽出した（以後、「症状名詞-句症状表現（候補）」と呼ぶ）。これにより、6,353 パタンの症状名詞-句症状表現（候補）を得た（表 5-3）。

3.2 抽出された句症状表現の評価

身体部位-句症状表現候補、症状名詞-句症状表現候補それぞれを Web テキスト上で出現したバイナリパタンの共起頻度降順に並べ、上位 1,000 までを対象に評価を行う（表 5-4）。これら 1,000 件中に適切な句症状表現が含まれる正解率を評価指標として採用した。

取得した句症状表現の比較には、患者表現辞書 [3]を用いた。患者表現辞書（正式版、2019/03/26）には、全 8,862 レコードが存在する。出現形には重複があるため、出現形の異なりを除外したところ、レコード数は 8,440 件となった。この中には「感覚異常」、「神経痛」などのような症状名詞の他に、「眼が乾燥する」のような本研究の対象としている＜名詞＋助詞＋述語＞パタンの句症状表現も登録されている。このような表現を同定するために、Juman 辞書を用いた MeCab による形態素解析を行い、品詞を付与した。各クエリの後方一致による検索とマッチング数を表 6 に示す。

表 6　患者表現辞書

検索クエリ	一致数
名詞_格助詞_(動詞|形容詞|形容動詞)$	1,847
名詞_格助詞_名詞_(動詞|形容詞|形容動詞)$	186
名詞_格助詞_(動詞|形容詞|形容動詞)_接尾辞$	585
名詞_格助詞_(動詞|形容詞|形容動詞)_助動詞$	7
合計/総エントリ数	2,625/8,440

4. 結果

機械学習で構築した身体部位名詞分類器と症状名詞リストが、本研究対象の名詞に対しても有効に適用できているかを確認するために、表 5-3 における句症状表現の名詞に対して、身体部位名詞か否か、もしくは症状名詞か否か、それら以外であればどのようなカテゴリに分類されるのかについて、全例サンプリングを行い評価した。同時に、症状テンプレートの表現の使われ方についても評価を行った。症状表現に用いられうる表現を「収集対象」として集計し、それ以外に対して可能な限りカテゴリ分類を行った（図 3）。

身体部位-句症状表現候補の名詞では、94.9%の精度で身体部位を判定できていた。また、症状名詞-句症状表現候補では、名詞リスト構築時の予測モデル精度よりやや低下しているものの、84.2%と高い精度で症状表現を判定出来ていた。

表 5-4 の句症状表現候補の判定結果を表 7 に示す。本手法は身体部位名詞と症状名詞のどちらの場合でも約 8 割で句症状表現を特定できる精度であることを確認した。また、確認した 1,578 表現中、1,376 例（87.2%）は患者表現辞書に収録されていない新規の表現であることを確認した。

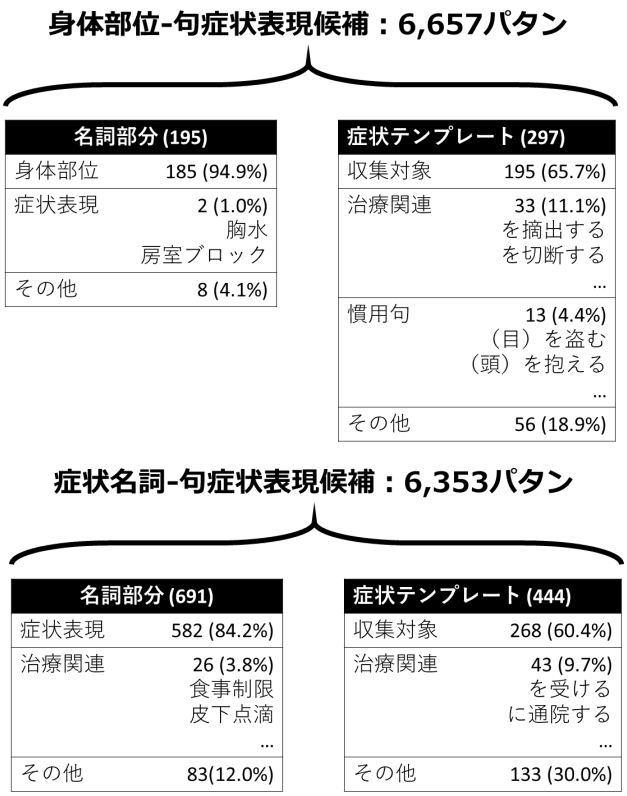

図 3　句症状表現候補の名詞・症状テンプレート分類

表 7　提案手法による句症状表現の抽出精度

	句症状表現		
	身体部位	症状名詞	身体部位名詞&症状名詞
正解率	77.7%	80.1%	78.9%
患者表現辞書にない表現	663/777 (85.3%)	713/801 (89.0%)	1,376/1,578 (87.2%)

患者表現辞書には存在しない句症状表現の具体例を表 8 に示す。

表 8　新規に獲得された句症状表現

身体部位−句症状表現	喉をやられる、歯が溶ける、腰が曲がる、歯を失う、歯がない、気管支が弱い、声を失う、頭がおかしい、顔につくる、肌が荒れる、目がかゆい、肌が弱い、膝をつく、…
症状名詞−句症状表現	発疹がひどい、言葉を発す、体調を崩す、嘔吐を繰り返す、記憶を失う、高熱が出る、出血が続く、頭痛がする、高熱を出す、傷が付く、BPSDが目立つ、…

5. 考察

本研究の提案手法により、高精度で新規の句症状表現を Web 文書に出現した表現パタンから取得することが出来た。

事前学習した BERT モデルのファインチューニングにより構築した身体部位名詞分類器は、抽出出来た名詞に関しては 94.9%と高い精度で身体部位名詞を特定できていた。しかしファインチューニング時の Recall が 78.3%と低いことが課題であった。原因精査のために development データにおける真陽性事例 (True Positive: TP)と偽陰性事例 (False Negative: FN)を確認した。その結果、一語として認識された筋肉名が真陽性事例に比べて偽陰性事例に多いことがわかった（TP:

腕_撓骨_筋、鎖骨_下_筋、腰_回旋_筋、上腕二頭筋、… / FN: 腰棘間筋、頭長筋、大円筋、体幹筋、…)。事前学習の段階で、日本語 Wikipedia コーパスではこれらの細かい解剖学的単語の関係性を学習するのに不十分であった可能性や、Juman 辞書に基づいた形態素解析結果をそのまま用いたことが原因と考えられた。これを解決するには、Byte Pair Encoding として WordPiece[15]などを導入してサブワード分割を適用する手法や、単語末尾の一文字の特徴を単語入力時に与えることで、「～筋」で終わる情報を BERT モデルに入力する手法が挙げられる。本研究では、このような詳細な解剖学的名称を使用した症状表現が Web 文書に頻出する可能性は低いと考えて、対象の名詞 41,717 語に構築した身体部位名詞分類器をそのまま適用し、汎用的に身体部位を特定できる結果が確認できた。しかし、一部の身体部位名詞が偽陰性として判定されている可能性については、負例と判定された約 35,000 表現を全て確認することが出来ていないため、否定は出来ない。

以前構築した症状表現抽出モデルでは、症状表現かどうかの基準として症状名詞以外にも、所見(検査含む)、疾患名も許容した。また、名詞単体では症状表現とは言えないが、他の表現との係り受けで症状表現になるものも対象にしていた。今回の結果では、この基準が広く句症状表現を取得できる一因となった。例えば、「高血圧になる」、「高血糖が続く」という表現とともに、「血圧が高い」、「血糖値が高い」が句症状表現として確認できる。他にも「体重が落ちる」、「気分が悪い」、「意識を失う」などが収集した表現の中に存在した。

治療に関連する句表現も本手法では同時に収集されてしまう(例:歯を抜く、腎臓を切除する、…)。1,000 事例中 9.2%と症状表現以外では最も多い割合で存在した。これらの多くは症状テンプレート部分に特徴的なものが多い(例:～を切除する、～を移植する、～を摘出する)。しかし、症状表現として使われるものも含まれるため、症状テンプレートのみの判定では精度の改善は難しい(例:「歯を抜く」は治療だが、「髪を抜く」は抜毛症の症状と言える)。句症状表現候補だけではなく、文脈内での使われ方で判定が出来る手法を今後検討する。

また、身体部位から句症状表現を抽出する方法の課題として、慣用句の存在が挙げられる。今回評価した 1,000 事例のうち、28 事例が慣用句表現であった。その一部を以下に示す。

息を引き取る、手に入れる、手を出す、目を盗む、頭がいる、目に遭う、頭を抱える、目にあう、耳にする、お腹に入れる、手がかかる、息を引きとる、腕を上げる、目を離す、肩を落とす、足を運ぶ、手にする、手を上げる、手に入る、目が行く、目にする、顔を出す、手を挙げる、手を付ける、手が止まる、腕を挙げる、手がつく、手をあげる、…

これらを機械による判定で除外するために、慣用句リストを入手して除外リストを作成するのも実現可能な手法と思われる。

本研究手法により取得した句症状表現は患者表現辞書にはない新規の表現が多く含まれていたが、一方で頻度上位 1,000(合計 2,000 表現)を確認した中には、患者表現辞書に存在する句症状表現 2,625 事例中、202 事例 (7.7%)しか重複は存在しなかった。その理由を検討するために、患者表現辞書に収録されている句症状表現を確認し、表 9 のように分類した。

症状を表すオノマトペの多くは、「～になる」、「～する」などの表現を伴う。本研究では「～になる」に伴うオノマトペは「ボロボロになる」が取得出来ていたが、「～する」に関しては「頭がする」のような表現に集約されているため症状と判断できなかった。このような症状のオノマトペを伴う動詞のパタンを認識することで、元の文に戻り該当する表現を獲得して症状表現を完成させることが可能になると考えている。

否定表現に関しては、今回の表現取得資源が WISDOM X のファクトイド QA データベースのインデックスであり、述語部分が原形に正規化されていたため取得対象外となっていた。今回取得した表現の否定形が Web でどの程度用いられているかを頻度で評価し、症状表現として妥当なものを抽出する手法も考えられるが、表 9 の例のように「見えない」だけではなく「見えにくい」と多様な接尾辞表現を取りうるため、このような否定表現パタンを準備することから始める必要がある。

患者表現辞書に収録されている表現の中には、本研究のように、使われやすい表現を組み合わせて句症状表現を生成する手法では、取得が困難な表現が存在する(表 9-4)。今後さらに症状表現を獲得していくには、患者表現辞書に出現するパタンを参考にした上で、実際に使用される患者症状表現を収集し、表現辞書の拡張を行うことを検討している。具体的には、看護記録を用いて患者の主観的情報(S: subjective)フィールドに記載された表現から患者症状表現の学習・評価用データを構築する手法や、初診患者問診システムに症状記入欄を設け、表現の収集を行うことなどを行う予定である。

表 9　患者表現辞書の句症状表現分類

1	オノマトペ	・喉がゼイゼイする ・踵がガサガサする
2	否定表現	・眼が見えない ・眼が見えにくい ・息が出来ない
3	動詞の名詞化による句症状表現	・飲み込むのが難しい
4	長文	・トイレに行く回数が多く寝ていても目が覚めてしまう
5	名詞・格助詞を 2 つずつ伴う表現	・両手足に力が入らない
6	症状名詞リストにない表現	・ゲロを吐く
7	検査値表現	・Zn の数値が上がりすぎ
8	確認範囲外に存在した句症状表現	・みぞおちが痛い(1,197 位) ・痙攣が起きる(1,250 位)

6. 結論

本研究では、名詞の特徴に注目して句症状表現を Web 文書から効率的に取得する手法を提案した。

事前学習済みの BERT モデルに自身で準備したデータを用いてファインチューニングを施すことで、高精度な単語分類器を構築することが出来た。

本手法により獲得した単語リストと症状テンプレートは、柔軟にお互いを組み合わせることで句症状表現を生成することが可能と思われる。その妥当性評価を Web 文書に出現したかどうか、その出現頻度はどのぐらいかを提示することで、症状表現である可能性が高いものを優先的に人手で選択し、評価することも可能になる。

本研究により、一般の患者が使う可能性の高い句症状表現が特定でき、症状名詞への紐付けもしくは疾患名との関連付けができれば、患者が入力する多様な症状表現に基づいて疾患名を推定することのできる診断支援システム構築の第

一歩になるものと期待する。

参考文献

1) 万病辞書. MEDNLP 医療言語処理グループ 万病辞書. 【http://sociocom.jp/~data/2018-manbyo/ (cited 2019-Aug-10)】
2) Comejisyo. ComeJisyo プロジェクト日本語トップページ – OSDN. [https://ja.osdn.net/projects/comedic/ (cited 2019-Aug-10)]
3) 患者表現辞書. Patient Disease Expression 患者症状表現辞書. 【http://sociocom.jp/~data/2019-pde/ (cited 2019-Aug-10)】
4) 患者症状抽出器. MEDNLP 医療言語処理グループ AEX. 【http://sociocom.jp/~data/2017-AEX/ (cited 2019-Sep-03)】
5) Mizuno J, Tanaka M, Ohtake K, et al. WISDOM X, DISAANA and D-SUMM: Large-scale NLP Systems for Analyzing Textual Big Data. COLING 2016 : 263-7.
6) Wada S, Iida R, Torisawa K, Takeda T, Manabe S. Extracting Symptom Names and Disease-Symptom Relationships from Web Texts Using a Multi-Column Convolutional Neural Network. In Proceedings of MEDINFO 2019 : 423-7.
7) 厚生労働省. 平成３０年版医師国家試験出題基準について. [https://www.mhlw.go.jp/stf/shingi2/0000128981.html (cited 2017-July-3)]
8) MEDIS. ICD10対応標準病名マスター. [https://www2.medis.or.jp/stdcd/byomei/ (cited 2017-July-10)]
9) 情報通信研究機構. ALAGIN 言語資源・音声資源サイト. [https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html (cited 2017-July-3)]
10) Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
11) Kazama J, Torisawa K. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. Proceeding of ACL-08 2008 : 407-15.
12) Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. arXiv : 1901.08746, 2019.
13) Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv : 1904.05342, 2019.
14) Kudo T, Yamamoto K, Matsumoto Y. Applying conditional random fields to Japanese morphological analysis. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing 2004 : 230–7.
15) Wu Y, Schuster U, Chen Z, et al.. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144, 2016.