医師国家試験自動解答プログラムの処方薬問題への拡張

Extension of Question-Answering Program to Automatically Answering the Medical Licensing Examination to Drug Related Questions

水口 達矢
Tatsuya Mizuguchi

慶應義塾大学理工学部
Faculty of Science and Technology, Keio University
mizuguchi@dna.bio.keio.ac.jp

伊藤 詩乃
Shino Ito

(同 上)
shino@dna.bio.keio.ac.jp

佐藤 健吾
Kengo Sato

(同 上)
satoken@bio.keio.ac.jp

榎原 康文
Yasubumi Sakakibara

(同 上)
yasu@bio.keio.ac.jp, <http://www.dna.bio.keio.ac.jp>

keywords: question-answering, medical licensing examination, medical diagnosis, Word2vec.

Summary

Medical diagnostic support system is an automatic support system that prevents doctors from unknowingly mis-interpreting medical results. However, it is not an easy task to automate the procedure with high accuracy. Our goal is to construct such a medical diagnostic support system that could improve the overall accuracy of medical diagnoses. As a pilot study, we built a program that automatically answers the medical licensing examination (MLE), in our previous study. MLE involves questions that require the users to pick answers such as disease names or drug names from multiple choices, given the patient information. In our previous study, the program was developed to answer only disease related questions, but we realized that the study will not be complete without deciding optimal drug for patients. For this reason, we attempt to expand this program to answer drug related questions in the current research. The major improvements include vectorizing the words and automizing the construction of rule base. By this, we prevented the tedious task of inputting drug information manually and now it is possible to avoid influences of inconsistent spelling and synonyms by vectorization of words. We managed to increase the accuracy of the previous study up to 56.1%.

1. はじめに

近年、日本では医療分野の情報化が進んできており、その例として電子カルテの普及があげられる [厚生 18]。電子カルテとは、従来、紙のみでしか存在しなかった患者の診察記録を医師がコンピュータ等を利用して診断と同時並行的に電子化し、データベースとして保存・活用するシステムである。電子カルテの導入により、同一院内に限らず、遠隔地の医療機関同士における情報共有が容易に可能となり、患者の移送・移管時等における正確・迅速な医療の継続と、症状の見落としなどの初歩的診断ミスを未然に防ぐといった効果が期待されている。このように、電子カルテは医療機関における効率の向上を目指して導入されたものであるが、処方薬の副作用予測 [White 13] や、薬剤と病気の因果推定 [Wright 10] など、2次利用としても広く用いられている。

このことから、電子カルテの利用に着目し、電子カルテ

に記載された患者の情報を適切に読み取り、診断することのできる診断支援システムの開発を目指している。そのパイロット研究として、電子カルテと形式が類似している医師国家試験における臨床実地問題を、コンピュータにより自動解答するシステムの構築がされた [伊藤 16, 川村 16]。医師国家試験における臨床実地問題とは、問題文において患者の様々な情報（年齢・性別・症状等）や、検査結果が与えられ、それらの情報をもとに、選択肢の中から最も相応しいものを選択するという問題である。先行研究 [伊藤 16, 川村 16] では臨床実地問題のうち、病名を診断する問題（病名診断問題）を自動解答するシステムが構築され、診断支援システムの土台を築いた。しかし、医師国家試験には病名を診断する問題以外にも、患者に処方すべき薬を選択する問題（処方薬問題）や、次に行うべき検査を選択する問題（検査問題）などが存在する。病名を診断するだけでなく、上記の問題を適切に

図1 第109回医師国家試験の構成 [井廻 15]

解答できる能力は支援システムにとって重要である。

したがって本研究では、これまで開発してきたシステムを拡張し、処方薬を選ぶ問題を解答できるシステムの構築を目指す。また、本研究では、「ルールベース構築の自動化」および「単語のベクトル化」を行うことでシステムの改良を図った。「ルールベースの構築の自動化」により、事前に薬品の情報をシステム内にマニュアルで入力しておく必要がなくなり、「単語のベクトル化」により、表記揺れや異表記の影響を回避できることが期待される。

2. 背景

2.1 エキスパートシステム

エキスパートシステムは専門家の意思決定能力を模したものであり、ある特定の範囲の知識について問題解答することを目的とする。本研究で構築するシステムは、医学知識を有したエキスパートシステムである。医療分野において有名なシステムとして、MYCINが知られている [Shortliffe 75]。知識ベースと推論エンジンという大きく2つの部分からなっており、推論エンジンが知識ベースを用いて推論を行い解答を出すものである。MYCINは500程度の規則からなる知識ベースを持っており、医師に対して質問をいくつかすることで、伝染性の血液疾患を診断し、確率の高い推論リストを出すことができる。医療分野におけるエキスパートシステムの開発は現在でも続いており、継続して研究結果が報告されている [William 17]。2017年には、医療データベースや医師の知見から知識ベースを作成し、小児の胸痛の原因を判定することのできるエキスパートシステム [Randa 17] が構築された。また、PlusExpert [Devraj 11] は穀物の病気を診断するエキスパートシステムであるが、ファジー理論を導入することで曖昧性を与えた知識ベースが使用されている。

2.2 医師国家試験臨床診断問題

医師国家試験の構成を図1に示す。医師国家試験は、医学知識を問う一般問題と、臨床における判断を問う臨床実地問題に分けられる。臨床実地問題はさらに、病名診断問題や処方薬問題などに分けることができる。

§1 先行研究

先行研究 [伊藤 16] において、医師国家試験における臨床実地問題を自動解答するシステムが構築された。臨

問診情報	32歳の男性。
	関節痛を主訴に来院した。本日未明から突然、右の第1趾に強い疼痛を感じた。意識は清明。
検査情報	身長160cm、体重70kg、体温37.0℃、尿素窒素28mg/dL、クレアチニン1.2mg/dL、尿酸8.2mg/dL、CRP 5.6mg/dL。
画像情報	足の写真を次に示す。
	現時点の治療薬として適切なのはどれか。

図2 臨床実地問題の構成

床実地問題とは、問題文において患者の様々な情報（年齢・性別・症状等）や、検査結果が与えられ、それらの情報をもとに、選択肢の中から最も相応しいものを選択するという問題である。この研究では臨床診断問題のうち、患者の情報から、疑われる病名を診断する問題に取り組み、これらの問題を解答できる病名診断システムが構築された。この病名診断システムは Demner-Fushman ら [Demner-Fushman 07] および Soldaini ら [Soldaini 15] の類似の研究と間接的ではあるが精度評価を行っており、どちらの研究よりも高い精度を示した。しかし、臨床実地問題には病名診断問題以外にも処方薬を選択する処方薬問題が存在し、診断支援システムにとって処方薬問題を解答する能力は病名診断能力と同様に重要である。以上のことから、本研究では病名診断システムを拡張し、処方薬問題を解答できるシステムの構築に取り組むことを本研究の目的とする。

§2 臨床実地問題の構成

臨床実地問題の問題文の例を図2に示す。問題文には、患者の「問診情報」、「検査情報」、「画像情報」が与えられる。

- 問診情報
患者の主訴、付随症状が示されている。
- 検査情報
患者の数値データおよび検査データが記されている。
- 画像情報
患部を中心に、CT スキャン等の画像が載せられている。

検査情報と画像情報は、すべての問題文中に含まれているわけではなく、必要に応じて記載されている。今回解答対象とする処方薬問題は臨床実地問題の1つであり、上記の形式で作成されている。

臨床実地問題は5肢択一の選択問題であり、5つの選択肢の中から、患者に適した選択肢を選択する。

§3 臨床実地問題の課題

臨床実地問題では、5つの選択肢に与えられた病名や処方薬の中から解答を選択する。そのため、臨床実地問題を分類問題として捉えると、病名や処方薬の種類だけ分類クラスが存在することになる。したがって、扱う問

図3 1クラスあたりの学習データ数

題数に対する分類クラスが非常に多くなり、機械学習を行う際に十分に学習されないという問題が生じる。処方薬問題を例に、30年分の試験問題155問中における、各処方薬の出現回数を図3に示す。ここから、過半数の処方薬が、一度しか試験問題に現れていないことがわかる。今回のような分類クラス数に対する学習データ数が極端に少ないケースにおいて高い精度を得るためには、事前知識の獲得や事前学習が有効である [Fei-Fei 06a, Fei-Fei 06b]。病名診断システムでは、事前知識として、病名の情報を格納したルールベースを事前に構築する手法が取られた。この際、ルールベースは医学教科書からマニュアルで構築されていた。本研究でも同様に事前知識として処方薬の情報を格納したルールベースを構築するが、公共データベースから情報を取得することにより、自動でルールベースを構築できるように変更を加える。また、本研究では事前学習として Word2vec [Mikolov 13] を用いた単語のベクトル化を行う。これにより、「嘔気」≒「吐き気」のような単語関係の定量化を行うことができる。

2.3 処方薬問題の課題

処方薬問題と病名診断問題の関係を図4に示す。処方薬を判断するためには、患者に考えられる病名を診断し、その上で病気に効く薬品を選択する必要がある。図4の例を用いると、まず問題文に記載された患者にみられる発熱や腎不全といった症状から、痛風発作や家族性地中海熱を診断し、その後、これらの病名と患者の症状とを合わせて総合的にコルヒチンが有効であると判断される。このように、処方薬問題を解答するシステムは、病名診断システムと統合することで構築することができる。一方で、処方薬は病気を患っていない来院者に処方されることもあり、そのため病名診断問題と統合しても解答ができない問題が存在する。例えば、抗不安薬や睡眠薬は健康な人にも処方されるが、この場合は病名診断を介さずに処方薬を選ぶ必要があり、病名診断システムと統合するアプローチでは適切に解答することができない。

そこで本研究では病名診断システムとの統合は行わず

図4 病名問題と処方薬問題の関係

に、直接処方すべき薬品を解答するアプローチを提案する。このアプローチでは、医学教科書や添付文書に記載されている薬品の効果効能に加え、薬品の身体への作用などの幅広い情報を収集することで、症状から直接処方薬を判断できるようにする。以上のことから、これまで手動で行われていたルールベース構築を自動化することで、より多くの情報を収集することができる環境を構築することは重要である。

2.4 電子カルテとの比較

臨床実地問題は電子カルテと同様の記述形式がとられており、非構造化データ（患者の問診等）、構造化データ（検査値等）、画像データが記載されている。したがって臨床実地問題から診断を行うことのできるシステムは、大きな修正を加えることなく電子カルテへの適用が可能である。一方で電子カルテには略語の利用が多いことが知られている [Liu 15] ため、これに対応するには略語を網羅したシソーラス辞書を作成する必要がある。また、この課題は電子カルテから Word2vec を学習させることでも対応できる [Liu 15]。

3. 方 法

3.1 解 答 対 象

解答対象として第80回から第109回までの医師国家試験臨床実地問題を用いた。この問題セットのうち、本研究では、問題文において患者の症状・検査結果などが示され、選択肢の中から適切な処方薬を解答する問題を処方薬問題とみなし、解答対象とした。この条件に当てはまる問題は、155問存在した。病名診断システムでは問題文に画像が含まれている問題を除外していたが、本研究では [川村 16] の手法を統合することで、画像付き問題も解答対象とした。

3.2 システムの構成

システムの概要を以下に示す。図5に示す通り、問題解答システムは、問題文を読み込み数値や重要な単語（キーワード）を抽出するプログラム、選択肢における情

図5 問題解答システムの構成

図6 問題解答システムの構成

報（ルール）を取得するプログラム、Word2vecにより単語の分散表現を構築するプログラム、以上の情報からスコア計算を行い、解答を出力するプログラム、学習を行うプログラムから構成されている。以降、それぞれにおける詳細を記載する。

3.3 キーワード抽出

臨床診断問題における問題文は「問診情報」、「検査情報」、「画像情報」の3つの情報に分割することができる。それぞれ異なるアプローチによってキーワードへの変換を行った（図6）。

§1 問診情報

問診情報では、文章を形態素解析器 MeCab[工藤 04]により形態素解析を行い、名詞と動詞の原形をキーワードとして抽出した。MeCabにはユーザー辞書として Come-Jisyo[相良 14]を使用し、医学用語に対応させたものを用いた。また、病名診断システム同様に CaboCha[工藤 02]による否定表現の検出を行い、否定語「ない」がかかっている単語はキーワードから除外した。例えば「頭痛がない」という文章の場合、否定語「ない」が「頭痛」にかかっているため、「頭痛」はキーワードとして抽出されないことになる。

§2 検査情報

検査情報は、形態素と数値の単語関係によって意味を持つため、先ほどのように形態素をそのままキーワードとして使用することができない。そこで、構造化を行った上で、その検査値が正常か異常であるかの判定を行った。構造化は、臨床実地問題における検査情報が、いくつかの規則にしたがって記述されていることを利用している。例えば「尿酸は8.2 mg/dLであった」という文章が与えられると、「尿酸：8.2 mg/dL」といった構造化した情報として取得することができる。この構造化された情報から、統計的指標である基準値によって正常か異常かの判断を行った。その後、異常と判断された場合には、これにより疑われる症状や病名を判定し、これらをキーワードとして取得した。これにより「尿酸：8.2 mg/dL」という情報は、基準値である7.0 mg/dLを超えているため「尿酸：高」と判断され、「痛風」や「発熱」といったキーワードに変換されることになる。基準値や疑われる症状などの情報は、ウェブサイトである『わかる pro』^{*1}と『これであなたも医の達人』^{*2}から取得した。

§3 画像情報

[川村 16]で提案されたアルゴリズムを使用して、画像情報をキーワードに変換した。そのアプローチについて記す。まず、患部の画像が問題文中に与えられると、これに最も近い画像をトレーニングデータ内から検索する。その後、検索された画像につけられたアノテーションをキーワードとして抽出した。具体的には、ORB アルゴリズム（Oriented FAST and Rotated BRIEF）[Rublee 11]によって、与えられた画像と、トレーニングデータ内のすべての画像において画像間類似度を計算し、最も類似度が高い画像の探索を行った。また、画像に対するアノテーションは医師国家試験過去問データベース^{*3}の解説文より取得した。この解説を問診情報同様に形態素解析により単語に分割することで、名詞と動詞の原形をキーワードとして抽出した。

3.4 ルールベースの作成

選択肢中には薬品名のみが与えられるため、その薬がどのような効果を持っているのか、どのように作用するのか等の情報を事前知識としてシステムに与える。本研究では選択肢に薬品名が与えられると自動で公共データベースに検索をかけ、必要な情報を取得するプログラムを作成した。使用するサイトは『おくすり110番』^{*4}と『LIFE SCIENCE DICTIONARY』^{*5}の2つである。『おくすり110番』では、薬品名を検索することで、その薬品における「概説」、「作用」、「効能」の3つの説明文を閲覧することができる。

*1 <http://kawaru.biz/>

*2 <http://tatujin.net/>

*3 <https://medu4.com/>

*4 <http://www.jah.ne.jp/kako/>

*5 <http://lsd-project.jp/cgi-bin/lsdproj/ejlookup04.pl>

図7 選択肢におけるルールの取得
(医薬品データベース：お薬110番より作成)

図8 同義語変換を用いたルールの取得
(同義語データベース：LIFE SCIENCE DICTIONARYより作成)

- 概説
薬品についての大まかな説明
- 作用
体内における働き、影響、効果が文章で記載
- 効能
対象の病名や病状、症状などが箇条書きで記載

これらの情報を取得することで、その薬品の特徴が得られる。本研究では、「概説」、「作用」、「効能」の情報をすべて取得し、形態素解析を行った。その後、名詞と動詞の原形のみを抽出し、行に薬品名、列に抽出した単語を持つ行列を薬品のルールとした（図7）。行列の要素には、3.8節に説明する機械学習のために、それぞれ取得時に重み $w=1$ を設けておく。例えば、薬品の説明文章中に「うつ病を改善する薬」という文が得られると、その薬品の行の要素として「 $w_{うつ病}=1$ 」「 $w_{改善}=1$ 」「 $w_{薬}=1$ 」が付与されることになる。ルールとして重要な単語は3.8節の学習において w が高くなるように学習されるため、説明文章中に出現する単語の回数や医学単語であるかにかかわらず、すべて $w=1$ としている。同様に、必要のない単語は w が小さくなるように学習されるため、否定表現の抽出も行っていない。このように作成された行列をルールベースと呼ぶ。

しかし、薬品名には同義語が多数存在し、問題文の表

記のまま『おくすり110番』で検索ができるとは限らない。例として、「オキミナス」と「ロキソニン」は同じ化合物を指しているが、『おくすり110番』には「ロキソニン」として登録されているため、「オキミナス」として検索しても情報が得られない。そのため選択肢として「オキミナス」が与えられた際には、「ロキソニン」に変換してから説明文章を検索しなければならない。『LIFE SCIENCE DICTIONARY』では、このような同義語（異表記）の情報が記載されている。「オキミナス」で検索すると、「ロキソニン」を含む同義語の一覧が得られる。したがって、選択肢に与えられる薬品名から直接『おくすり110番』で説明文書を取得できない場合には、『LIFE SCIENCE DICTIONARY』から同義語一覧を取得し、それらを順に検索していくことで対応する（図8）。

3.5 分散表現の構築

3.3節で問題文から抽出されたキーワードに、3.4節で取得した各薬品のルールが適用されることにより、薬品ごとにスコアがつけられる。どのようにルールを適用するかによって、全く異なるスコア、異なる解答につながる。簡単な適用手法として、[伊藤16]で用いられた、問題文のキーワードと処方薬のルール中の単語との間で一致している単語の数をスコアとする方法がある。この手法の欠点として、以下のことが考えられる。

- 単語単位で比較される
例)「頭が痛い」→「頭」+「痛い」≠「頭痛」
- 表記揺れに弱い
例)「コリアー徴候」≠「コリエー徴候」
- 同義語に弱い
例)「嘔気」≠「吐き気」
- 似た表現と、全く異なる表現が同じ扱い
例) (「熱」と「高熱」)=0, (「熱」と「骨折」)=0

これらの問題を解決する方法の1つとして、単語同士の関係性を何らかの方法で定量化することがあげられる。本研究では、この定量化のために、Word2vecを使用する。Word2vecであらかじめ、単語の分散表現（ベクトル表現）を作成しておき、これを参照することで問題文のキーワードや、薬品のルールの分散表現を獲得する。PythonライブラリであるGensimのWord2vecを使用し、あらかじめMeCabにより分かち書きをしておいた日本語Wikipediaのダンプデータ^{*6}をコーパスとし、分散表現モデルを作成した。次元数およびwindow sizeをハイパーパラメータとし、その他のパラメータはデフォルトを指定した。学習モデルとして、Skip-gram、階層的Softmaxを選択した。以上により、213,460単語の分散表現が作成された。

^{*6} <https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2>

図9 スコア行列の作成

3.6 スコア行列の作成

問題文のキーワードと、薬品を説明するルール中に現れる単語の関係を数値で表し、表の形式でまとめたものをスコア行列と呼ぶ。本研究では、3.5節で獲得した分散表現によって、問題文のキーワードと、薬品のルール中の単語をすべてベクトル化し、コサイン類似度を用いることによって問題文と薬品の説明文章との関係性を評価する。スコア行列は薬品ごとに用意され、図9のように、キーワードとルール中の単語の行列であり、単語間のコサイン類似度に、ルールの重みがかけられたものを要素とした。コサイン類似度の特徴として、同じベクトル同士、つまり同じ単語同士の類似度は1になる。したがって、コサイン類似度でスコア行列を作成することは、前節で述べた、キーワードとルール中で一致している単語の数をスコアとする方法を、内包していることになる。また、関節炎と関節痛のような似た単語がコサイン類似度が大きく、吐き気と関節痛のような概念の大きく異なる単語はコサイン類似度が小さくなる傾向があるため、より詳細に単語関係を定量化することが可能である。

3.7 解答の判定

§1 薬品スコアの計算

前節で作成したスコア行列をもとに、選択肢である薬品にスコアをつける。

本研究では、薬品ごとのスコア行列に対して各キーワードの行中の要素の最大値をキーワードスコアとし、このキーワードスコアの合計をその薬品のスコアとした。これは問題文中のキーワードの集合と選択肢の薬品のルールの集合の類似度を、各キーワードと最も類似度が高い薬品のルール中の単語とのスコアを用いて計算していると捉えることができる(図10)。

問題文のキーワードを K 、薬品 d のルールの集合を R_d 、あるルール r の重みを w_r とすると、スコア計算は以下の式で与えられる。ここで、 $\cos(r, k)$ は r と k のコサイン類似度である。

$$\text{score}_{\max} = \sum_{k \in K} \max_{r \in R_d} w_r \cos(r, k)$$

図10 スコアの計算

このスコアは5つの選択肢すべてに対して計算される。

§2 解答の選択

1つの問題において、選択肢の薬品ごとにスコア行列が作成され、そこからスコアが求められる。1問あたり5つの選択肢(薬品)のスコアが求まるため、このスコアが最大のものを解答とした。複数選択問題では、スコアを降順にソートし、上位の薬品から解答とした。

3.8 学習

システムが出力した解答から誤差を計算し、誤差が小さくなるようにルールベースの重みを学習する。ルールの重みの初期値はすべて1である。臨床診断問題では5つの選択肢にスコアをつけ、高いものから解答としていくため、問題ごとに5クラス分類問題であるとしみなすことができる。また、1つの問題に対して解答が2つ以上の場合が存在する。以上のことから、本研究では誤差関数としてシグモイド交差エントロピーを用いて学習を行った。最適化手法はAdamを採用した。

4. 実験と考察

4.1 処方薬問題における正答率

処方薬問題155問を対象とした正答率を表1に示す。ハイパーパラメータ(学習率、Word2vecの次元数およびwindow size)の最適化にはベイズ最適化[Snoek 12]を使用し、過学習を防ぐために、バリデーションデータセットの正答率が最も高くなる学習回数をテストデータの学習回数とし、早期終了を行った。

表1において、(A)~(C)はWord2vecによる単語のベクトル化を行わず、ルールとキーワードに共通して存在する単語の数を選択肢のスコアする既存手法である[伊藤16]。(A)では問診情報のみを使用し、これに加えて(B)では検査情報を、(C)では検査情報と画像情報から得られたキーワードを使用して解答している。これに対し、(D)~(F)はベクトル化を用いた手法である。ここから、問診のみを使用した場合と比べて、検査、画像の情報を加えることによって正答率が向上していることがわ

表1 本手法における正答率

ベクトル化	使用する情報	正解率 (%)
なし	A. 問診	37.4
	B. 問診+検査	42.6
	C. 問診+検査+画像	42.6
あり	D. 問診	50.3
	E. 問診+検査	55.5
	F. 問診+検査+画像	56.1

表2 正答率の比較

手法	正解率 (%)
G.Siamese-LSTM	31.3
H.QA-LSTM with max pooling	40.0

かる。さらに、(D)~(F) 手法ではいずれもベクトル化を行わなかった手法に比べ高い正答率が得られた。この要因として、単語のベクトル化により表記揺れや異表記に対応できるようになったことが考えられる。

また、本手法における (B) の正答率が先行研究 [伊藤 16] で示された病名診断システムの正答率よりも低い原因としては、2.3 節で述べたように、処方薬問題が病名診断問題と比較して難易度が高いことが考えられる。

4.2 比較手法

本手法との比較として、深層学習を用いた2つの手法を実装した。これらの手法による正答率を表2に示す。

● Siamese-LSTM [Mueller 16]

文章間類似度を測るタスクにおいて、高い精度を持つ LSTM (Long short-term memory) を用いた手法である。文章 A と文章 B を重みを共有する2つの LSTM に入力し、出力された2つのベクトルの距離を類似度スコアとする。

● QA-LSTM with max pooling [Tan 15]

質問応答において、LSTM を使用して質問文と応答文の距離を求める。max pooling によって局所的な情報を取り出すことができる。2つの LSTM の重みは [Feng 15] に基づき共有させる。

上記の手法では、2つの LSTM に対して、一方の LSTM に問題文を、もう一方の LSTM に薬品の説明文章を入力し、この2つ文章が正解のペアである場合には出力スコアが1に、そうでない場合には-1に近づくように学習を行った。問題を解答する際には1つの問題文に対し、選択肢ごとのスコアをそれぞれ計算し、スコアが高いものを解答としている。本稿では、提案手法同様に Wikipedia のダンプデータを用いて Word2vec の事前学習を行い、その結果を LSTM の入力とした。また、Word2vec および LSTM のハイパーパラメータはベイズ最適化によって決定した。

これらの手法は汎用性が高く、様々な文章間類似度や質問応答のタスクにおいて高精度の結果を示しているが、

表3 キーワード抽出の評価 (代表例)

	対応できたもの	未対応のもの
問診情報	胃部不快感	Gram / 染色
検査情報	総蛋白	MCV
画像情報	十二指腸潰瘍	肺炎

表4 学習後のルールベースの重み

薬品	ルール	重み (w)
メトロニダゾール	トリコモナス	1.26
インフリキシマブ	一般	1.15
コルヒチン	ベーチェット病	1.11
ARB	高血圧症	1.10
オキシトシン	末期	1.09
スピロラクトン	水分	0.734
カナマイシン	肝性脳症	0.734
メトロニダゾール	関連	0.737
セフェム系抗生物質	感性	0.737
硝酸薬	急性期	0.738

本研究においては、提案した手法 (D) がこれらの正答率を上回った。この要因として2.2.3 節で述べたように、臨床実地問題では、扱う問題数に対して分類クラス数 (処方薬数) が非常に多いということがあげられる。(G),(H) で用いられている Siamese-Network は多クラス分類問題を2クラス分類問題として扱えること、組み合わせによりデータ数を増幅できることから学習サンプルが少ないデータにおいても高い精度を期待できるモデルである [Koch 15] が、今回のような極端なケースでは、パラメータの多い LSTM では十分に学習することができなかったことが考えられる。

4.3 キーワードの評価

キーワード抽出において、1問あたり、問診情報から平均32.4単語、検査情報から平均20単語、画像情報から平均6単語のキーワードが抽出された。適切にキーワード抽出が行えたものと、改善が必要なものについていくつかの代表例を表3に示す。問診情報では「胃部不快感」を1つのキーワードとして抽出することができた一方で、「Gram 染色体」を「Gram」「染色体」の2つのキーワードとして抽出してしまった。この課題を解決するには MeCab のユーザー辞書を拡張することが必要である。検査情報では、血圧や総蛋白など48項目においてキーワード抽出を行うことができた。しかし MCV ($1,000 \times Ht \div RBC$) のように記載されている検査値からさらに計算が必要な項目に関しては、今のシステムでは対応できていない。画像情報に関しては、キーポイント検出において十二指腸潰瘍などの局所的な特徴を捉えることができた。検出できなかった特徴としては、肺炎のような、ORB で捉えられない大きな特徴などがあげられる。

表 5 第 80 回問題 D38 におけるスコア

	キーワード	薬品のルール (重み)	キーワードのスコア	薬品のスコア
ペニシリン G (問題の正解)	発熱	発熱 (1.0)	1.00	6.59
	咳嗽	呼吸器 (0.98)	0.71	
	細菌	菌種 (0.77)	0.54	
メチシリン (システムの解答)	細菌	細菌 (1.0)	1.00	10.00
	上昇	低下 (0.9)	0.72	
	球菌	菌種 (1.0)	0.64	

表 6 第 80 回問題 D3 におけるスコア

	キーワード	薬品のルール (重み)	キーワードのスコア	薬品のスコア
アミトリプチリン (問題の正解)	食欲不振	うつ状態 (1.0)	0.761	7.85
	不眠	うつ状態 (1.0)	0.753	
	症状	片頭痛 (1.0)	0.711	
ジアゼパム (システムの解答)	症状	症候 (1.0)	0.774	8.66
	食欲不振	自律神経失調症 (1.0)	0.735	
	不眠	身体 (1.0)	0.720	

4.4 ルールベースの評価

本システムにおいて 316 (薬品数) × 2031 (ルール数) のルールベースが構築され、すべての問題の選択肢中の、全薬品に対応することができた。475 (病名数) × 907 (ルール数) である病名診断システムと比較すると本手法で構築したルールベースの方がルール数が多くなっている。また、学習後のルールベースについて、重みが大きくなったルールと小さくなったルールをそれぞれ上位 5 つずつ表 4 に示す。ここから、重みが大きくなったルールの方に病名 (トリコモナス、ペーチェット病) や症状 (高血圧症) が多く含まれていることがわかる。一方で肝性脳症の重みが小さくなっているが、肝性脳症はカナマイシンの応用利用であり、一般的な利用法ではないためであると考えられる。

4.5 エラー分析

本節では、システムが正解できなかった問題の詳細および改善策について述べる。

第 80 回問題 D38 『大酒家。5 日前から 38℃ 台の発熱と咳嗽とがあった。2 日前から左側胸部に激痛があり、体温が 40℃ に上昇したので来院した。白血球 14,300。左胸水貯留を認める。胸水は膿性、粘稠で腐敗臭が強く、多数の Gram 陽性球菌と多核白血球がみられるが、通常の培養で、24 時間後までは細菌の発育がみられない。』

この問題の正解はペニシリン G (ベータラクタム系抗生物質) であるが、システムはメチシリン (ペニシリン系抗生物質) と誤って解答してしまった。表 5 に各薬品のキーワードスコアのうち、値が大きかったものを示す。ここから、メチシリンのスコアが高くなった原因として、「(体温の) 上昇」に対して「(免疫機能の) 低下」のルールの

スコアが高い (0.72) ことがあげられる。「上昇」と「低下」は反対の意味を持つ単語であるが、どちらも「増減」を表す単語という同じ性質を持っているため、Word2vec におけるコサイン類似度が近くなってしまっている。また、「Gram 陽性細菌」が「Gram / 陽性 / 球菌」と形態素解析され、この「球菌」という単語がメチシリンのスコアの向上を引き起こしてしまっている。したがって Gram 陽性細菌」を 1 つのキーワードとして取り出すことができるよう、MeCab のユーザー辞書を更新する必要がある。

第 80 回問題 D3 『3 ヶ月前から食欲不振、胃部不快感および早朝の嘔気があった。この間に体重が 10kg 減少した。精密検査の結果、特に異常なく、胃神経症として治療を受けていたが、症状の改善がみられなかった。不眠が続き、意欲も集中力もなく仕事が思うようにできず悩んでいる。』

この問題の解答はアミトリプチリン (抗うつ薬) であるが、システムはジアゼパム (抗不安薬) と誤って解答してしまった。表 6 に各薬品のキーワードスコアのうち、値が大きかったものを示す。不正解となった原因として、重みの学習が行われていないこと (すべて 1.0 のまま)、および症状や症候のような特徴的でない単語の影響が考えられる。前者は学習サンプルを拡充することで改善が期待される。後者は、TF-IDF 等によって、ある薬品に特徴的なルールの重みを大きな値で、どの薬品にも共通しているルールの重みを小さな値で初期化することが必要である。

5. 結 論

先行研究において構築された病名診断問題自動解答システムを拡張し、処方薬を選択する問題を自動解答するシステムを構築した。システムは大きく分けて 4 つのプ

プログラムに分けることができ、それぞれに対して以下の改良を加えた。

- (1) 選択肢における情報（ルール）を取得するプログラム
これまで手動で構築されていたルールベースの自動構築を可能にした。
- (2) 各選択肢にスコアをつけ、適切な解答を出力するプログラム
キーワードとルールに共通している単語の数をスコアとしていた従来の手法から、Word2vec によるベクトル化を用いた計算に変更したことにより、表記揺れや異表記に対応できるようになり、学習の向上に繋がった。
- (3) 問題文から患者の情報を読み取るプログラム
画像からキーワードを取り出すプログラム [川村 16] を取り入れることによって、画像付き問題が解答可能になった。
- (4) 機械学習によりシステムを更新するプログラム
固定増分誤り訂正法から Adam による最適化に変更した。また、ベイズ最適化によりハイパーパラメータの探索を行った。

新たに提案したシステムを用いて、第 80 回～第 109 回医師国家試験の処方薬問題 155 問について解答した結果、56.1 % の正答率が得られ、従来の手法を上回った。

今後の課題として、文章を単語に分割し単語単位で扱っていることから、語順や文脈を無視していることがあげられる。文章を単語として扱っていることにより、情報量が少なくなることは明確である。そのため、より精度の高いシステムを構築するためには、語順や文脈を加味したスコア計算手法が必要である。また、本研究のマイルストーンは医師国家試験に合格できるシステムを構築することであり、そのためには 70 % 近くの正答率を必要とする。さらなる正答率の向上のためには、上記の課題を解決するほか、臨床実地問題の練習問題や実際の電子カルテなどの利用により学習サンプルを拡充することが有用である。

本研究では処方薬問題を対象としたが、検査問題を解答する能力も診断支援システムには重要である。臨床検査は、患者に疑われる病気の確定診断として行われるため、病名診断システムと統合することで検査問題を解答することができる。一方で、検査問題では選択肢が文章で与えられることがあるため（例：血中活性型ビタミン D の測定）、3・4 節と同様に、選択肢をデータベースで検索可能な形に変換するプログラムの構築が必要である。

謝 辞

本研究は、文部科学省「革新的イノベーション創出プログラム (COI STREAM)」，慶應義塾大学 COI トライアル拠点のプロジェクトの一環として行われた。また本研究は、「私立大学戦略的研究基盤形成支援事業」の支援

を受けている。

◇ 参 考 文 献 ◇

- [相良 14] 相良 かおる：ComeJisyo の紹介と医療情報に含まれる誤字調査，情報知識学会誌，Vol. 24, No. 2, pp. 204–209 (2014)
- [Demner-Fushman 07] Demner-Fushman, D. and Lin, J.: Answering clinical questions with knowledge-based and statistical techniques, *Computational Linguistics*, Vol. 33, pp. 63–103 (2007)
- [Devraj 11] Devraj, B. and Jain, R.: PulsExpert: An expert system for the diagnosis and control of diseases in pulse crops, *Expert Systems with Applications*, Vol. 38, pp. 11463–11471 (2011)
- [Fei-Fei 06a] Fei-Fei, L.: Knowledge transfer in learning to recognize visual objects classes, in *International Conference on Development and Learning* (2006)
- [Fei-Fei 06b] Fei-Fei, L., Fergus, R., and Perona, P.: One-shot learning of object categories, *Institute of Electrical and Electronics Engineers TPAMI*, Vol. 28, pp. 594–611 (2006)
- [Feng 15] Feng, M., Xiang, B., Glass, R., Wang, L., and Zhou, B.: Applying deep learning to answer selection: A study and an open task, in *Proceedings of Institute of Electrical and Electronics Engineers ASRU Workshop* (2015)
- [井廻 15] 井廻 道夫：医師国家試験改善検討部会報告書 (2015), <http://www.mhlw.go.jp/file/05-Shingikai-10803000-Iseikyoku-Ijika/0000079678.pdf> [Accessed 10 Jan. 2018]
- [伊藤 16] 伊藤 詩乃, 田中 佑岳, 狩野 芳伸, 榊原 康文：医師国家試験を自動解答するプログラムの構築，人工知能学会学会誌，Vol. 31, No. 6, pp. 1–9 (2016)
- [川村 16] 川村 健史, 伊藤 詩乃, 榊原 康文：検査画像情報を取り入れた医師国家試験自動解答プログラムの構築，第 2 回医用人工知能研究会 (SIG-AIMED) (2016)
- [Koch 15] Koch, G., Zemel, R., and Salakhutdinov, R.: Siamese neural networks for one-shot image recognition, in *International Conference on Machine Learning Deep Learning Workshop* (2015)
- [厚生 18] 厚生労働省：医療情報化の体制整備の普及を推進すること (2018), https://www.mhlw.go.jp/file/05-Shingikai-12601000-Seisakutoukatsukan-Sanjikanshitsu_Shakaihoshoutantou/0000091465.pdf [Accessed 29 July. 2018]
- [工藤 02] 工藤 拓, 松本 裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，Vol. 43, pp. 1834–1842 (2002)
- [工藤 04] 工藤 拓, 山本 薫, 松本 裕治：Conditional random fields を用いた日本語形態素解析，情報処理学会論文誌，Vol. 161, pp. 89–96 (2004)
- [Liu 15] Liu, Y.: Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion, in *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)* (2015)
- [Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed representations of words and phrases and their compositionality, in *Advances in Neural Information Processing Systems*, Vol. 26, pp. 3111–3119 (2013)
- [Mueller 16] Mueller, J. and Thyagarajan, A.: Siamese recurrent architectures for learning sentence similarity, in *Proceedings of Association for the Advancement of Artificial Intelligence* (2016)
- [Randa 17] Randa, K.: Rule based system for chest pain in infants and children, *Indonesian Journal of Electronics and Instrumentation Systems*, Vol. 1, pp. 138–148 (2017)
- [Rublee 11] Rublee, K. K., Vincent Rabaud and Bradski, G.: ORB: An efficient alternative to SIFT or SURF, Vol. 13, pp. 2564–2571 (2011)
- [Shortliffe 75] Shortliffe, E. H. and Davis, R. e. a.: Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system, *Computers and Biomedical Research*, Vol. 8, No. 4, pp. 303–320 (1975)
- [Snoek 12] Snoek, J., Larochelle, H., and Adams, R.: Practical bayesian optimization of machine learning algorithms, in *Neural Information Processing Systems* (2012)
- [Soldaini 15] Soldaini, L.: Retrieving medical literature for clinical

decision support, in *European Conference on Information Retrieval* (2015)

[Tan 15] Tan, M., Santos, C., Xiang, B., and Zhou, B.: Lstm-based deep learning models for non-factoid answer selection, *arXiv preprint arXiv:1511.04108*. (2015)

[White 13] White, T. N., NH, S., RB, A., and E, H.: Web-scale pharmacovigilance: Listening to signals from the crowd, *Journal of American Medical Informatics Association*, Vol. 20, pp. 404–408 (2013)

[William 17] William, W.: Trends in expert system development, *International Journal of Engineering and Information Systems*, Vol. 1, pp. 85–96 (2017)

[Wright 10] Wright, C. E. and FL, M.: An automated technique for identifying associations between medications, laboratory results and problems, *Journal of Biomedical Informatics*, Vol. 43, pp. 891–901 (2010)

〔担当委員：清田 陽司〕

2018 年 5 月 28 日 受理

著 者 紹 介

水口 達矢

2017 年慶應義塾大学理工学部生命情報学科卒業。同年 4 月より同大学院理工学研究科基礎理工学専攻生命システム情報専修修士課程に在籍。

伊藤 詩乃

2015 年慶應義塾大学理工学部生命情報学科卒業。2017 年同大学院理工学研究科基礎理工学専攻生命システム専修修士課程修了。人工知能を用いた医師国家試験自動解答プログラムの開発を行った。同年 4 月よりアビームコンサルティング株式会社所属。

佐藤 健吾(正会員)

2003 年慶應義塾大学大学院理工学研究科博士課程修了。博士(工学)。慶應義塾大学理工学部助手を経て、2006 年よりバイオ産業情報化コンソーシアム研究員、産業技術総合研究所生命情報科学研究センター共同研究員、2009 年より東京大学大学院新領域創成科学研究科特任講師、2011 年より慶應義塾大学理工学部生命情報学科専任講師。バイオインフォマティクス、自然言語処理、統計的機械学習などに興味を持つ。国際計算生物学会、日本バイオインフォマティクス学会、情報処理学会、各会員。

榊原 康文(正会員)

1985 年東京工業大学大学院理工学研究科修士課程修了。(株)富士通研究所情報社会科学研究所研究員、米国カリフォルニア大学サンタクルーズ校客員研究員を経て、1996 年東京電機大学理工学部助教授、2002 年慶應義塾大学理工学部生命情報学科助教授、2005 年同教授、現在に至る。博士(理学)。バイオインフォマティクスを専門とし、オミックスデータからデータマイニングによる情報解析、知識発見、創薬などの研究に従事。