共起単語スコアによる医師国家試験の解答に向けて

○二宮仁志[†]　松木萌[†]　大北剛[‡]　井上創造[‡]
[†]九州工業大学　[‡]

Toward Answers National Examination for Medical Practitioners by Word Co-occurrence Score

Hitoshi Ninomiya, Moe Matsuki, and Okita Tsuyoshi, and Sozo Inoue
[†] Kyushu Instiute of Technology

Abstract : *In this paper, we aim at automatically answering National Examination for Medical Practitioners. We consider a pair of words between questions and choices of correct answers often appears in the knowledge source text, and introduced word co-occurrence based on PMI. As a result of solving some problems of the doctor national exam using the word co-occurrence score, we obtained the accuracy of 29.85% compared to 20% in case of random answer.*

1　はじめに

医師国家試験は日本の国家資格である医師免許を取得するための国家試験であり，試験内容は各回 500 問の選択問題からなる．本研究は，医師国家試験の自動解答を行なうシステムを構築することを目的とする．

医師国家試験の自動解答システムとして，伊藤らは病名判定ルールを用いて解答するアプローチを取った[1]．伊藤らは，医学用語シソーラス 39 万語を用いて各病気に対する症状や特徴などの単位と数値を抽出しておき，これらを特徴とした病名判定ルールを構築した．この病名判定ルールを用いて，選択問題の解答を自動選択する．医師国家試験に対処するものは我々の知る限り他にないが，大学入試を自動解答するシステムは本研究と目的を同じくする．世界史の共通一次試験の選択問題を自動解答システムにおいて，小林らは単語共起スコアを用いたシステムを構築した[2]．本稿では，医師国家試験の選択問題と解答を用いて，単語共起スコアを用いて最も確からしい解答を選択するシステムを構築した．伊藤らは医学用語シソーラス約３９万語を用いたが，我々は日本語 Wikipedia のデータ約 4.2 億語のみという資源の限られた設定で行なった．

単語共起を測定するために PMI (Pointwise Mutual Information：自己相互情報量)を用いる研究は，1990 年の Church の研究に遡る[3]．[3] は，看護師の単語が出現する頻度は，文章内で医師という単語を含んでいる頻度に関連し，この統計量は PMI によって測定できることが示された．その後，PMI は文章中の特定のパターンを抽出するための道具[4]，因果関係のルールを抽出するための道具[5]などとして幅広く用いられている．

本稿では，医師国家試験の一部の問題について PMI を参考とした単語共起スコアを用いた解答について説明し，結果と考察について述べる．

2　医師国家試験

医師国家試験は, A 問題から I 問題まで 9 つのセクションが存在し，全 500 問を解答するものである．ほとんどが選択問題であるが，中には計算問題も存在する．

本稿では，第 100 回から第 111 回の医師国家試験の過去問題を，文献 [6]から Web スクレイピングにより取得した．[6]の過去問データベースには，質問文と選択肢，正解の解答が記載されている．また, いくつかの問題に対しては解説文と問題の分野（「子宮腺筋症」や「血管肉腫」など）が記述されていることもある．

3. 方法

本節では, 選択式問題における質問文と正答の選択肢の間の単語のペアを用いた解答手法を述べる．

3.1 システム概要

問題解答のシステム概要を以下の図 1 に示す．

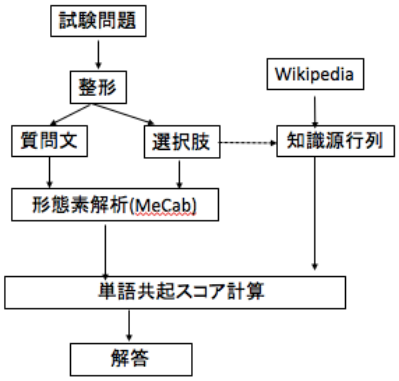

図 1.システム概要

全体の流れとしては，まず試験問題を整形し質問文と選択肢に分け，それぞれを形態素解析し問

題文と選択肢の間の単語ペアを作成する．次に，分類された選択肢を元にWikipediaから知識源行列を作成する．この単語ペアと知識源行列を用いて，自己相互情報量(PMI)に基づく共起確率を計算する．以下ではこの共起確率を**単語共起スコア**と呼ぶ．

3.2 整形

医師国家試験の問題を行列化し，分析しやすい形にする．今回用いた過去問は，2 章で述べたようにWeb スクレイピングによって HTML 形式で取得する．取得した HTML 形式から質問文，解答，選択肢(a,b,c,d,e)，解説，テーマを構造分析により抽出し，行列とする．その結果，6033×9 の列となった．

3.3 知識源の抽出

本稿では，「診断」を解答する問題に焦点を当てる．その理由や問題の分類については 3.4 節で述べる．これらの問題は，いくつかの「症状」から推測される「診断」を問う問題である．この問題に対して，選択肢の「症状」と質問文の「症状」を比較する解答方法が考えられる．このためには，知識源に求められる情報として，「診断」に対する「症状」の記載があるものを抽出することが重要である．

日本語 Wikipedia では，1 単語につき 1 ページが割り当てられ，さらにページ内で「歴史」「定義」などのカテゴリが存在する．そのため，診断名に対応する Wikipedia ページが存在すれば，「症状」に関連する記述カテゴリとして記載されており，抽出が容易であると考える．

選択肢の診断名を行，Wikipedia ページ内のカテゴリを列とする行列を生成し，これを**知識源行列**とする（図 3）．

また知識源行列は，高速化のため単語共起スコアを計算する際に各問題に必要な箇所のみを抽出する．

3.4 質問解析

問題を解く際に，質問応答において重要である問題のカテゴリの種類は以下の 2 つであると考えられる．

1. 問題は何を聞いているのか
2. 肯定文か否定文か

図 2.知識源行列の作成方法(問題の選択肢から取得した単語をキーワードとし，Wikipedia のページを検索する．そして選択肢の単語を行，Wikipedia のページ内のカテゴリを列とした知識源行列を生成する)

1 つ目の意味は，解答する属性で分類することを考えている．例えば「この患者の診断はどれか.」という問題は，「診断」を解答することを要求されており，「症状」を答えてはならない．この「診断」にあたる単語を属性として，分類することを考える．本稿では，「診断」を解答する問題に焦点を当てる．

「診断」を解答する問題の抽出方法は，以下の流れで行う質問文の最後の 1 行を抽出する．

1. 抽出された文章の係り受け解析を行い，
2. 最後の単語(*word_0*)とその単語に係る単語(*word_1*)を抽出し，
3. 抽出された *word_0* が「どれか」かつ，*word_1* に「診断」が含まれている問題を抽出する．

係り受け解析には係り受け解析ツール CaboCha[8]を用いる．これにより，6033 問中 347 問の問題を抽出することができた．

2 つ目の「否定文か肯定文か」の問題例をあげると，肯定文が「適切な診断はどれか.」の場合に対して，否定文は，「適切でない診断はどれか.」である．分類方法は先ほどの抽出方法 1 で抽出した各問題の 1 行に「ない」もしくは「誤って」が含まれる質問文を否定文と定義する．その結果，530 問が否定文として分類された．

本稿では，肯定文である問題に焦点を当てる．

3.5 形態素解析

3.3 節の問題の分類で抽出された問題の質問文とそれに対応する選択肢を，形態素解析ツール MeCab[7]を使い形態素解析を行い，それぞれを形態素に分けた．形態素に分けた後，形態素ごとの「品詞」と「品詞細分類 1」をある条件の元に表層形を取り出し，各質問文とそれに対応する選択

肢ごとに単語の配列を作成し，スコア計算に用いた．

追加の辞書として ComeJisyo[9]を使用した．ComeJisyo は，医療施設で使われる医療用語を集めた辞書である．

3.6 単語共起スコアの計算

3.6.1 単語ペア

単語共起スコアを計算する際に，選択肢ごとに単語ペアを作成し，単語共起スコアの計算に用いた．

質問文と選択肢を形態素解析し，質問文と選択肢ごとに単語の配列を作成し，その質問文と選択肢ごとに単語ペアを作成した．この単語ペアを単語共起スコアに用いる．

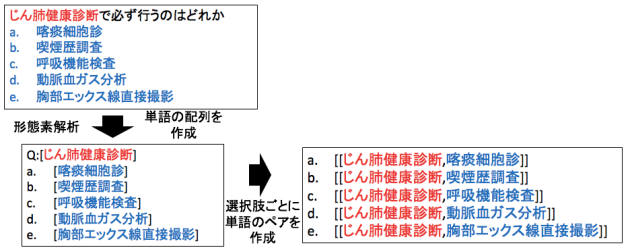

図 3.単語ペアの作成（質問文と選択肢を形態素解析し，質問文と選択肢ごとに単語の配列を作成し，その質問文と選択肢ごとに単語ペアを作成）

3.6.2 PMI(自己相互情報量)

PMI とは，ある 2 つの単語間の強さを共起確率で計算するものである．ある二つの単語，word1，word2 の PMI は，以下の式で計算できる．

$$PMI(word1, word2) = \log_2 \frac{P(word1, word2)}{P(word1)P(word2)}$$

$P(word1)$とは，文章における単語$word1$の出現確率，$P(word1, word2)$とは，文章中における単語$word1$と$word2$の共起確率を表している．

3.6.3 単語共起スコア計算

3.6.1 項で作成した単語ペアに対して下記の式で計算していく．

$$Score = \frac{1}{|S|} \sum_{(q_i, a_j) \in S} \log_2 \frac{P(q_i, a_j) + C}{P(q_i)P(a_j)}$$

ここで，単語q_iを質問文の単語,単語ajを選択肢の単語，S はq_iとa_jのペアの集合，|S|は S の要素数，$P(q_i, a_j)$は，知識源の文章数N中でのq_iとa_jの共起確率，$P(q_i)$と$P(a_j)$は，知識源の文章数N中での単語q_iと単語a_jの出現確率である．C は定数で，共起確率が存在しない場合の為に，C を加算した．スコア計算では，選択肢ごとにスコアの偏りが出ないようにするために，平均を取った．

3.6.3 選択肢の正誤判定

選択肢の正誤判定は，単語共起スコアが高い選択肢を正文である確率が高いとして選択した．

4. 評価方法と結果

以下の 2 点を評価する．1 つ目は単語共起スコアで解答した精度の評価として，単語共起スコア計算での解答とランダム解答の正答率の比較をした．2 つ目は，知識源を変えた単語共起スコアで解答した精度の評価に，知識源の絞り込み無しの正答率の比較をした．

4.1 共起スコアによる評価

第 100～第 111 回医師国家試験の問題のうち問題の分類「診断」で「肯定文」であるもの 272 問について解答を行なった．知識源は，3.4 節で分類した問題を元に作成したもので，今回は，「診断」の問題を解答するため，知識源行列の行に「診断」もしくは「症状」を含む列のみに絞り使用した．3.5 節で述べた単語ペアを作成するための，条件としては，「品詞」を「名詞」，「品詞細分類 1」を「一般」，「サ変接続」，「接尾」，「固有名詞」とした．また，今回は，単語共起で計算していくため，数値については考慮していない．

単語共起スコアが計算できた問題が 272 問中 168 問であり，計算できなかったものは 4 問であった．その 168 問中での単語共起スコアを用いた解答とランダム解答での比較を表 1 に示す．

	正答率
単語共起スコア	29.85%
ランダム	20 %

表 1. 単語共起スコア解答とランダム解答の比較

この結果より，ランダム解答より良い正答率より良い結果が得られた．

4.2 知識源の評価

知識源を絞って計算したものと絞らずに計算したものを比較し，結果を表 2 に示す．

	正答率
絞り込み有り	29.85%
絞り込みなし	28.73 %

表 2. 知識源の絞り込み有りと絞り込み無しの比較

この結果より，絞り込みをしていない知識源での解答より絞り込みした知識源での解答の方が，良い結果が得られた．

5. 考察

本章では，単語共起計算できなかった問題，正答できなかった問題についての原因追求，また知識源の絞り込み有無での解答についても分析を行い，現状と今後の課題について述べる．

5.1 単語共起スコアが計算できなかった問題

単語共起スコアが計算できなかった問題 4 問に

着目した．この 4 問において，選択肢の単語が抽出できていない選択肢が存在した．抽出できなかった理由として以下が考えられる．

(1)5.1 で述べた単語の抽出条件の漏れが存在していたこと．

(2)選択肢の間に空白が存在している為，二つの単語として扱ってしまい，抽出条件を満たさなかったこと．

(1)は，確認したところ抽出条件に足りなかったものが，”名詞”+”副詞可能”と”名詞”+” 形容動詞語幹”が条件として足りていなかった．これを新しく抽出条件に追加しスコア計算をした．結果としては，解答できた問題が，268 問から 271 問と向上したが正答率は,29.52%となり 0.32%低下した．

(2)の具体例としては，“が　ん”という選択肢が“が”と“ん”に形態素解析されてしまった．この選択肢の問題の他の選択肢をみてみると,”梅　毒“のように空白が存在しており，形態素解析をした場合，”梅”と“毒“と二つに，分かれてしまい全く違う意味の単語としてスコア計算をしてしまうことになる．このような選択肢を正しく抽出することは今後の課題である．

5.2 単語共起スコアを計算できた問題の中で正答できなかった問題

単語共起スコアを計算できた問題の中で正答できなかった問題に着目してみた. 正答できなかった理由として以下の原因があった．

(1) 正解である選択肢の知識源が存在していないもの．

(2) 選択肢の単語が形態素解析によって誤った別れ方をしているもの．

(1)は，選択肢の Wikipedia の題名と一致する Wikipedia のページが存在しなかった為である. 例えば，選択肢”Behcet 病”と題名が一致する Wikipedia のページは存在しない. しかし,”Behcet 病”を翻訳した”ベーチェット病”が題名となる Wikipedia のページは存在する. また,”肺結核” と題名が一致する Wikipedia のページは存在しない. しかし，”肺結核”という単語が”結核”というページの中の項目として存在している．

このように，翻訳した単語や，同義語で一致するページを取得すること．また，他のページの中の項目での存在する場合，階層構造を考慮し，取得することで, 知識源のない選択肢を減らすことが可能である．

(2)の具体例としては, ”Behcet 病”が” Behcet +病”, ”Eisenmenger 症候群”が”Eisenmenger + 症候群”, のように誤った別れをしているものがあった. ”病”と”症候群”は，医療文章の中で他の医療単語と共起しやすい為, 単語共起スコアに影響することが考えられる．”病”と”症候群”を単語抽出されないようにスコア計算をすると 29.85%から 30.22%となり，0.37%精度が向上した．

これらのことを考慮すると, 形態素解析をする際に，新しく辞書に登録しておくなどをし，誤った単語抽出を防ぐことは，今後の課題である．

6. ま と め

医師国家試験の「診断」,「肯定文」の問題を対象にし, 質問文の単語と選択肢の単語間の共起性を単語共起スコアで計算する手法を示し, 特定の問題において 29.85%の精度が得られた. 今後の課題として，知識源の不足の改善，単語抽出条件の改善，解答する問題の増加が挙げられる．

参考文献

[1] 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, 人工知能学会論文誌, Vol. 31, No. 6 (2014).

[2] 宮下洋 石井愛 小林実央 星野力: センター試験『世界史 B』文の正誤判定問題ソルバー, 言語処理学会 (2016)

[3] Kenneth Ward Church and Patrick Hanks (March 1990). "Word association norms, mutual information, and lexicography". Comput. Linguist. 16 (1): 22–29.

[4] Minimally Supervised Event Causality Identification, Quang Xuan Do, Yee Seng Chan, Dan Roth,EMNLP 2011.

[5] Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, an Varga, Jong-Hoon Oh, Yutaka Kidawara, 2014.

[6]医師国家試験過去問データベース: https://medu4.com

[7] MeCab : http://taku910.github.io/mecab/

[8] CaboCha: https://taku910.github.io/cabocha/

[9] ComeJisyo: https://ja.osdn.net/projects/comedic/