6V-04

中医学のための単語埋め込みに基づく情報検索手法の研究

太田遥人† 関隆志‡ 高橋晶子† 力武克彰†
仙台高等専門学校† フジ虎ノ門整形外科病院‡

1. 研究背景

近年，中医学は補完医療としての需要が高まっている．中医学の診断では，問診等から得られる患者の症状や状態から，病態を示す「証」を特定することにより， 治療方針を決定する． しかし，証の特定には証と症状を結び付ける多くの経験と知識が必要となり，診療経験の少ない医師の診断が困難となっている．従って，診療経験の少ない医師への診断を支援する仕組みが必要である．

診療経験の少ない医師への診断支援として，中医学文献の情報を基に症状から証を検索できるシステムを構築することが有用である[1]． しかし，中医学文献には表記ゆれや同義語が数多く存在し，検索システムを構築する際に支障となっている．近年，分散表現を用いることにより，同義語の解消に効果があることが知られている[2]．

従って，分散表現を利用し，中医学文献に対する情報検索手法を確立することができれば，医師の診断の支援を行うことが期待できる．

2. 研究目的

中医学文献に対する情報検索手法を確立させることを目的とし、「証を含む分散表現の獲得手法」と，「分散表現を用いた情報検索手法」の2つの手法を提案し，その有意性を検証する． また，医師の診断の支援を実現することを目的とし，情報検索システムの構築を行う．

3. 中医学文献に対する情報検索手法

分散表現とは，単語を高次元のベクトルで表現したモデルのことである．分散表現を用いることにより，単語の意味的な類似度をベクトルにより定量的に計算することができる．そのため，検索システムに応用可能であると期待できる．分散表現を用いて中医学文献から証の検索を行うためには，検索対象となる証を単語として扱うことのできる「分散表現のモデル」が必要となる．しかし，証が含まれた分散表現のモデルは過去に作成された例がない．そのため，証を含む分散表現のモデルを獲得することができれば，証と症状の単語間の意味的な類似度を分散表現上で比較することにより，検索を行うことが可能になる．

3.1. 証を含む分散表現の獲得手法

中医学文献を入手しデータ化することは困難であり，ニューラルネットワークによる分散表現の学習を行う際，精度が期待できるほどのテキストデータを集めることは難しいという問題がある． そのため， 既に学習済のモデルを利用して，証の分散表現を生成する手法を提案する．

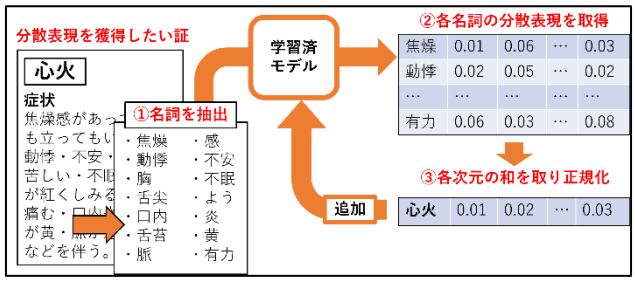

図 1. 証の分散表現の獲得手法

提案する分散表現の作成手順を「図 1」に示す．

① 分散表現を獲得したい証の症状が記述された文章から，名詞のみを抽出する．
② 学習済モデルを用いて各名詞の分散表現を取得する．
③ 各名詞の分散表現の和を取り正規化し， 証のベクトルとして学習済モデルに追加する．

以上の手法で，少ない文献データでも証の分散表現を獲得することができる．

3.2. 分散表現を用いた情報検索手法

分散表現では，意味の近い単語が近いベクトルとして表現され，ベクトルの近さであるコサイン類似度を計算することにより単語の類似度を得ることができる．本提案手法では，入力症状のベクトルに対し証のベクトルのコサイン類似度が高い順にランク付けすることにより検索を行う．コサイン類似度は， 入力症状のベクトルを $\boldsymbol{q}$, 証のベクトルを $\boldsymbol{d}$ として表すと式(1)のように定義される．

$$sim(q,d) = \cos(\boldsymbol{q},\boldsymbol{d}) = \frac{\boldsymbol{q}^T\boldsymbol{d}}{\|\boldsymbol{q}\|\|\boldsymbol{d}\|} \quad (1)$$

また，入力症状が複数の単語や文章であった場合，入力を形態素解析器で単語ごとに分解した後，その単語群のベクトルの重心を計算することにより， 入力症状のベクトルを得る．定義式を式 2 に示す．ここで， Q は入力文を単語ごとに分解した単語群である．

$$\bar{Q} = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{q_i}{\|q_i\|} \quad (2)$$

4. 情報検索システムの構築

本研究では，医師の診断の支援を実現するために， 提案手法を用いた検索システムを構築する．医師のデバイスに依存しないよう， Web アプリケーションとして構築を行う．

4.1. プロトタイプの構築

検索システムのプロトタイプの構築を行った．検索システムのプロトタイプを図 2 に示す．検索フォームに検索をしたい症状を入力すると，本提案手法により検索された証の情報がコサイン類似度の高い順に表示される仕様となっている．現在， 東洋医学専門医による検証により UI・表示情報の改善を行っている．

5. 検索システムの動作実験

本検索手法の動作を確認するために，症状から証の検索を行う実験を行った． 入力する症状は「発熱」とした．ランキング精度を確認するため，「発熱」と関係する証であるか否かの正誤の確認を行った．正誤判定は，東洋医学専門医の検証により行った． 実験結果を表 1 に示す． 表 1 は，検索結果全 75 件中， 閲覧される可能性の高い上位 10 件を抽出したものである．

今回の実験結果では， 上位 10 件中 7 件の正解の証を抽出することができた．しかし，不正解である証も見受けられた(順位 2 の水飲阻滞など)．理由として，これらの証の症状には，「発熱」と直接は記述されていないものの，「吐き気」「めまい」といった，発熱とコサイン類似度が近い単語が多く含まれていたことが分かった．この結果から，分散表現による検索は、コサイン類似度に基づく類義語を基に有用な情報を抽出できる可能性がある反面、不要な情報も類似度が高く出てしまう場合があることが確認できた。

今回の実験は「発熱」という症状に限った検索結果であり、全体的にどれほど表記ゆれ・類義語に対して効果があるかといった定量的な検索の精度に関しては測定できていない．そのため，提案手法の定量的な精度評価を行う必要があると考えられる． また， 検索システムについても，医師のフィードバックの元 UI や表示情報の改善を行っていく必要がある．

6. おわりに

本研究は，中医学において医師の診断を支援するために，証に関する情報検索システムを構築することを目的とし，分散表現を用いた情報検索手法の提案，検索システムの構築を行うものである．

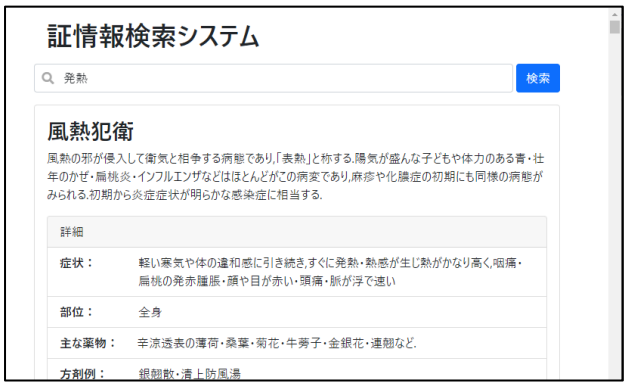

図 2. 検索システムのプロトタイプ

表 1. 「発熱」による検索結果上位 10 件

順位	証の名称	正誤
1	風熱犯衛	○
2	水飲阻滞	×
3	風熱犯衛兼裏実	○
4	陰暑	○
5	表実証	○
6	風熱頭痛	×
7	痰熱上擾	×
8	風寒襲表兼陰血虚損	○
9	暑熱傷気	○
10	風熱犯衛兼陰虚	○

今後の予定としては，検索手法の有意性を示すために，検索精度の評価を行っていく．具体的には，ベースとなる学習済モデルによる検索精度の差，従来の手法である文字列一致をベースとした手法との検索精度の差を比較する．

参考文献

[1] Ryo Nakagawa et al., “Design of a Diagnostic Support Method Utilizing Interrogation Information in Traditional Chinese Medicine”, The 35th International Conference on Advanced Information Networking and Applications, Proceedings AINA 2021, Vol 3, LNNS 227, 2021
[2] 田口雄哉, 田森秀明, 人見雄太, 西鳥羽二郎, 菊田洸, ” 同義語を考慮した日本語の単 語分散表現の学習, 情報処理学会研究報告”, Vol.2017-NL-233 No.17, 2017.

Information retrieval system based on word embedding for Traditional Chinese medicine
† National Institute of Technology, Sendai College
‡ Fuji Toranomon Orthopedic Hospital