6V-03

中国伝統医学における文書類似度による証間因果関係推定

齋藤　陸†　　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，高齢化の進展に伴って健康寿命の延伸を図ることが求められており[1]，高齢者の病気予防や健康維持が求められている．中国伝統医学(中医学)は疾病の発生を未然に防ぐ「未病先防」の概念などにより西洋医学を補完する補完医療の一つとして注目されている．中医学では患者の心身の状態を表す「証」を診断し，証に基づいて治療が行われる．証の間には，ある証が原因となって他の証が引き起こされるといった因果関係が存在し，医師は診断や治療の際に証間の因果関係を熟慮する必要がある．しかし，証間の因果関係に関する知識は体系的にまとまっておらず，複数の文献や診療経験に基づいて総合的に判断する必要があり経験の少ない医師の負担となっている．

2. 研究目的

本研究は，中医学の診断に有用な証間の因果関係を文献から自動的に推定することを目的をとする．そのため，中医学文献を用いた文書類似度によって証間の因果関係を抽出する手法を提案し，その妥当性を検証する．

3. 文書類似度による証間因果関係抽出

因果関係の抽出を行う文献としては，各証について証の特徴が書かれた説明文とその証を引き起こす別の証(病因)についての記述がそれぞれされているものを用いる．

心陽暴脱証	
説明文	心陽暴脱証とは，心陽が衰えて竭き，陽気が暴脱(急に脱ける)した証候である．
病因	寒邪が心陽をひどく傷付けたり，痰や瘀血が心竅を塞いだり，心陽虚が進行して発生する．

図 1 中医学文献における証の記述例

因果関係のある証間には，証の説明文と病因の記述の間に何らかの類似性があると考えられるため，文書比較によって，病因の証を特定することができると考えられる．そこで本手法では，ある証の病因の記述と他全ての証の説明文を比較し，類似度の高いものを病因の証として抽出することで，証間の因果関係を抽出する．

この際，各説明文/病因の記述文を特徴付ける文書ベクトルを生成し，文書ベクトル同士の類似度を算出することで文書比較を行う．また，文書ベクトル同士の比較にはコサイン類似度を用いる．

3.1 文書ベクトルの生成

文書ベクトルを生成する手法としては Sparse Composite Document Vectors (SCDV)[2]を用いる．SCDV は，単語ベクトル空間をガウス混合モデルと逆文書頻度(idf)の値によって意味の近い単語同士が近いベクトル成分を持つよう修正する手法である．SCDV を用いることで，類義語や表記ゆれのある単語同士が近いベクトル成分を持つようになり，単純に単語ベクトルの平均を文書ベクトルとするよりも各単語の意味的特徴を反映したベクトルの生成が可能となる．SCDV による文章ベクトル生成の流れを図 2 に示す．

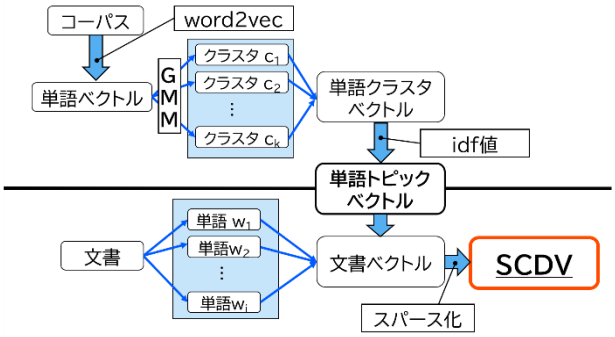

図 2 SCDV による文書ベクトル生成の流れ

4. 因果関係抽出実験

提案手法によって中医学文献に直接記載されている因果関係を抽出できるか確認するため，因果関係抽出実験を行う．

対象とする中医学文献は「全訳　中医診断学」[3]とする．対象文献において直接因果関係が記載

Estimation of Causal Relationships between Patterns Using text similarity in Traditional Chinese Medicine
Riku Saito† Takashi Seki‡ Akiko Takahashi† Yoshiaki Rikitake†
†National Institute of Technology, Sendai College
‡Fuji Toranomon Orthopedic Hospital

されている関係を手動で抽出し,原因→結果の因果関係のある証ペアを 19 件抽出し，正解データとする.

4.1 実験手順

実験手順を以下に示す.

Step1. 前処理（ストップワードの除去等）
Step2. 文章ベクトルの生成
Step3. 文書比較による証間因果関係の抽出
Step4. 評価

また，証間の因果関係抽出に適した文書ベクトル化手法を検討するため，Step2 において SCDV に加えて，SCDV とアプローチの異なる下記 2 種類のベクトル化手法によるベクトル生成を行い，結果を比較する.

- **TF-IDF:** 単語の出現頻度(TF)と逆文書頻度(IDF)から各単語の重要度を求めベクトルの成分とする手法
- **doc2vec**[4]**:** 文書から語順を反映したベクトル表現を直接獲得する手法

4.2 出力順位による評価

正解データの結果にあたる証について提案手法で因果関係の抽出を行い，類似度の降順に並べた際何位に原因にあたる証が出力されるかで評価を行う.

評価指標として MRR を導入する．MRR は，初めて適合文書が出現した順位の逆数を全検索対象について平均した検索評価指標で，以下の式で表される．適合項目が上位にあるほど 1.0 に近い値をとる.

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i}$$

N: **検索対象数**, r_i:**適合項目の順位**

4.3 実験結果と考察

各ベクトル化手法と MRR を表 1 に示す.

表 1 各ベクトル化手法と MRR

手法	TF-IDF	doc2vec	SCDV
MRR	0.534	0.480	0.590

TF-IDF と SCDV において MRR が 0.5 より大きくなっており，抽出結果上位 2 件以内に病因の証が出現していることが分かる．このことから，証間の因果関係に文書比較は有効であることが確認できる．また，中医学文献に直接記載された関係の抽出においては，SCDV が最も良い抽出精度を示した．また，TF-IDF によって類似度が 0 と算出された証についても SCDV では類似度が算出できていることを確認し，同義語や表記ゆれのある文書にも対応していることが分かった.

5. 東洋医学専門医による因果関係の評価

本手法によって中医学の診断に有用な証間因果関係が抽出できるか評価するため，提案手法によって病因と推定された証について東洋医学専門医による評価を行う.

5.1 評価方法

中医学文献中の各証について提案手法で病因の証の推定を行い，出力の上位 5 証について東洋医学専門医が関係の有無を判定することで評価する．本評価では，文書ベクトル化手法として SCDV を用いた場合について東洋医学専門医に評価を依頼する．評価指標としては MRR を用い，関係があると判定された証の出力順位に基づいて算出する.

5.2 評価結果と考察

評価結果から MRR を算出したところ，MRR は 0.68 となった．また，表 2 に関係があると評価された証の数の分布を示す.

表 2 関係のある証の分布

○の数	0	1	2	3	4	5
該当数	18	17	15	26	23	16

MRR が 0.68 となっていることから，提案手法による出力の上位 1.47 件に因果関係のある証が出力されていることが分かった．このことより，提案手法によって証間の因果関係が抽出できていることが分かる.

6. おわりに

本稿では，中医学の診断に有効な証間因果関係を文献から自動的に推定することを目的とし，文書類似度による証間因果関係抽出手法を提案し，その妥当性を確認した.

現在，東洋医学専門医に TF-IDF についても評価を依頼しており，それを基に SCDV を用いた因果関係抽出結果との比較を行っていく.

参考文献

[1] 厚生労働省,“健康寿命延伸プラン”, https://www.mhlw.go.jp/content/12601000/000514142.pdf (オンライン, 2021-12-24)
[2] Dheeraj Mekala et al.,” SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations”, Proc of EMNLP，pp．659-669，2017
[3] 浅野周，”全訳中医診断学”, たにぐち書店，2017
[4] Quoc V. Le, Tomas Mikolov, ”Distributed Representations of Sentences and Documents”, Proc of the 31st International Conference on Machine Learning, pp. 1188-1196, 2014