E-003

# 中国伝統医学における文献を用いた証間因果関係の抽出
# Extracting Pattern's Causal Relationships Using Text in Traditional Chinese Medicine

齋藤 陸 [1]　関 隆志 [2]　高橋 晶子 [1,3]　力武克彰 [1]
Riku Saito　Takashi Seki　Akiko Takahashi　Yoshiaki Rikitake

## 1. 研究背景

近年，伝統医学が補完医療として注目されており，特に中国伝統医学（中医学）は 2019 年に国際疾病分類第 11 版[1]の伝統医学分類に追加されているなど，関心が高まっている．

中医学では患者の心身の状態を表す「証」を診断し，証に基づいて治療や方剤の処方が行われる．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，ある証が進行すると他の証になるといった因果関係が存在し，診断や治療の際に証間の因果関係を考慮する必要がある．しかし，証の因果関係に関する知識は膨大な中医学文献を読み解くことでしか得ることができず，経験の少ない医師の負担となっている．

## 2. 研究目的

本研究では，中医学に熟達していない医師に向けて証間の因果関係を自動的に提示することによる診断の支援を目的とする．

証間の因果関係は中医学文献中に具体的に記述されている関係もあれば，文献の各証の説明文や証を引き起こす原因(病因)に関する記述から判断する必要がある関係も存在する．そこで本研究では，中医学文献中の各証の説明文や病因の記述を用いて自然言語処理によって診断支援に有効な証間の因果関係を自動的に抽出する手法を提案する．

## 3. 証間因果関係抽出手法

証間の因果関係の抽出手法として，本研究ではベクトル表現を用いた抽出手法を提案する．中医学文献では，証の説明文はその証を特徴付けるように記述されており，証の病因の記述はその証を引き起こす別の証に関する内容が記述されている．そこで，証の説明文からベクトル表現を得ることで証を特徴付けるベクトル(説明ベクトル)が獲得でき，同様に，証の病因の記述からその証を引き起こす別の証についてのベクトル(病因ベクトル)が獲得できると考えられる．

ベクトル空間上である証の病因ベクトルと距離の近い別の証の説明ベクトルが存在する時，それらの証間にはある証とその証を引き起こす証という関係がある可能性が高いと考えられる．ここから，ある証によって他の証が引き起こされるというような証間の因果関係を自動的に抽出することができる．ベクトル表現を用いることで，文中の表現の差異や複数の証についての記述があった場合でもベクトル空間上での距離が近ければ証間の因果関係を抽出することができる．

本手法では，初めに中医学文献から証の説明ベクトルと病因ベクトルを生成し，次に証の病因ベクトルとその他の証の説明ベクトル間の距離を求め，距離の近い証同士を因果関係がある証として抽出する．

### 3.1 各証の説明/病因ベクトルの生成

中医学文献における各証の説明文と病因の記述から文書分散表現(文書ベクトル)を生成し，説明ベクトルと病因ベクトルとする．

文書から文書ベクトルを獲得する手法として Sparse Composite Document Vectors (SCDV)[2]を用いる．SCDV は，word2vec[3]などで作成された単語ベクトル空間をガウス混合モデルと idf 値により修正することで意味の近い単語同士が近いベクトル成分を持つようにする手法である．SCDV を用いることで，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 1 に示す．

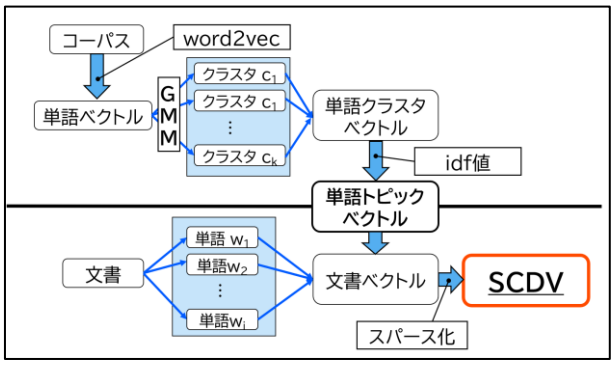


図 1　SCDV による文書ベクトルの生成

### 3.2 証間の因果関係の抽出

3.1 で生成した各証の説明ベクトルと病因ベクトルを用いて証間の因果関係を抽出する．初めに，ある証の病因ベクトルと他の全ての証の説明ベクトルとの距離を求める．次に距離の近い上位 N 件の証についてある証の原因となっている証であると見なし，因果関係として抽出する．他の全ての証についても同様の抽出処理を行い，因果関係を抽出する．ベクトル間の距離尺度としてはコサイン類似度を用いる．

## 4. 因果関係抽出・評価実験

本手法によって中医学で実際に知られている証間の因果関係を抽出できるかを評価するため，中医学文献である「全訳中医診断学」[4]を対象として提案手法を適用して証

1 仙台高等専門学校 National Institute of Technology, Sendai College
2 フジ虎ノ門整形外科病院 Fuji Toranomon Orthopedic Hospital
3 東北大学 Tohoku University

間の因果関係抽出を行い，抽出された上位 N 件の関係について評価する実験を行った.

### 4.1 実験の評価

対象とした中医学文献の病因の項目に「A 証が B 証に進行する」のように直接因果関係が記載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作成する.作成した証ペアの結果にあたる証について,提案手法によって証ペアの原因にあたる証が抽出できるかを評価する.ここで作成した証ペアの結果にあたる証から原因にあたる証を抽出できた件数を「ヒット数」とし,抽出件数 N を変化させた際のヒット数の比較を行う.

### 4.2 実験手順

実験は以下に示す手順で行う.

1) 各証の説明文と病因の記述から名詞を抽出し，表現の統一等の前処理を適用
2) 説明ベクトル/病因ベクトルの生成
   ① word2vec モデルの作成と学習
   中医学文献をコーパスとして python のライブラリである gensim[5]を用いて word2vec モデルを作成し，学習を行う.
   ② word2vec による単語ベクトルの生成
   ③ SCDV による単語ベクトルの修正と文書ベクトルの生成
3) 提案手法に基づく証間の因果関係の抽出
4) 抽出された因果関係の評価

また，実験手順の 2)をベクトル表現の生成方法の異なる以下の 2 手法と置換した場合と提案手法を用いた場合とでヒット数の比較を行う.

A) TF-IDF によりベクトル化を行う手法 (分散表現を用いない)
B) gensim の doc2vec によりベクトル化を行う手法 (文書から直接分散表現を獲得)

### 4.3 実験結果

実験結果として，各手法におけるヒット数を図 2 に示す.この際，中医学文献から抽出された証数は 115 証,評価対象の因果関係のある証ペアは 19 件であった．今回 N としては 1，12 (全証数の約 10%)，35 (全証数の約 30%)を選択した.

## 5. 考察

TF-IDF による手法 A では，N=1，12 の際には提案手法である SCDV の結果よりも良い結果が得られた．これは，TF-IDF では説明文と病因の記述の間で完全に一致する単語が現れた際に類似度が大きく上昇することに起因すると考えられる．しかし，TF-IDF では説明文と病因の記述の間に完全一致する単語が存在しない場合に類似度が 0 になってしまうという問題があり,因果関係の抽出が困難となる.

doc2vec による手法 B では，同じく N の増加に伴ってヒット数は増加しているものの他の手法と比較するとヒット数が低い結果となった.

SCDV を用いた提案手法では， N=35 の際に手法 A のヒット率よりも良い結果が得られた．これは完全に一致する

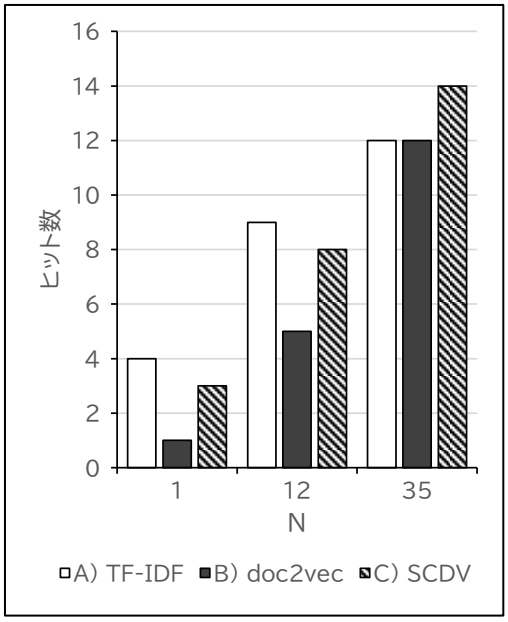


図 2　各手法におけるヒット数

単語が存在しない場合でも分散表現を用いることで類似度が算出できるようになり,手法 A で類似度が 0 になってしまうケースでも類似度を算出できることにより関係を抽出できたためであると考えられる.

## 6. おわりに

本研究では，中医学の診断支援に有効な因果関係を自然言語処理によって自動的に抽出することを目的とし，中医学文献中にある証の説明と病因の記述を用いて作成した各証の説明ベクトルと病因ベクトルのコサイン類似度を用いて証間の因果関係を抽出する手法を提案した.

結果として，抽出件数 N が小さい場合は TF-IDF を用いた手法の精度が高かったが，N を増加させていくと提案手法の精度が TF-IDF の精度を上回った．また，分散表現を用いたことで説明文と病因の記述中に完全に一致する単語がなくとも類似度を算出できることが分かった.

提案手法では，単語の生起情報のみを用いて証のベクトル化を行ったが，文脈や文章の構造も因果関係の表現に深く関わっているため，今後はそれらを考慮したモデルを手法に取り入れることを含め検討を行っていく．また，提案手法で抽出した因果関係について中医学医師に評価を依頼し，更なる手法の改善や精度の向上を行っていく.

### 参考文献

[1] WHO, “ICD-11 for Mortality and Morbidity Statistics” (オンライン), 入手先 <https://icd.who.int/browse11/l-m/en> (参照 2021-06-15)
[2] Dheeraj Mekala et al.,” SCDV : Sparse Composite Document Vectors using soft clustering over distributional representations”, Proc of EMNLP，pp．659-669，2017
[3] Tomas Mikolov et al., “Efficient Estimation of Word Representations in Vector Space” (オンライン), 入手先 <https://arxiv.org/pdf/1301.3781.pdf> (参照 2021-06-15)
[4] 浅野周，”全訳中医診断学”, たにぐち書店，2017
[5] Radim Řehůřek, ”Gensim: Topic modelling for humans”, <https://radimrehurek.com/gensim/> (アクセス日 2021-06-15)