O-015

中医学の診断を支援する表現の異なりを考慮した分散表現獲得手法の検討

Distributed Representation Acquisition Method Considering Variation of Representation to Support Diagnosis in Chinese Medicine

髙橋 唯*1 Takahashi Yui　関隆志*2 Seki Takashi　力武 克彰*1 Rikitake Yoshiaki　髙橋 晶子*1,3 Takahashi Akiko

*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

1. はじめに

中医学は，伝統医学の一部として国際疾病分類第 11 版に新たに導入されるなど補完医療としての需要が高まっている[1]．中医学は望診，聞診，切診，問診の四診により心身の状態を示す「証」を特定し治療を行う．特に，問診は患者の自覚症状を直接把握することができるため，患者の証の特定に有効とされる．問診で用いられる問診票[2]は，医師によっては，A3 用紙 8 枚に 1000 項目を超え，同じ症状であっても患者によって記入する表現が異なる．そのため，証の特定には表現の異なりを考慮する必要がある．しかし，同義語をはじめとする表現の異なりは患者や文献によって多様である．そのため，表現の多様性を考慮した証の特定が求められる．本研究では，日本語の文章を用いた分散表現の獲得やファインチューニングを行うことで，患者や文献による表現の多様性を考慮した文章の比較ができる中医学に特化した分散表現獲得手法を提案している．本稿では中医学文献をコーパスとして利用した分散表現の獲得について述べる．

2. 関連研究と提案

2.1 関連研究

中医学の診断支援に関する研究として，患者の問診内容や診断された証などの医療データを用いて機械学習を行い，患者の問診内容から証を特定する研究[3]がある．この研究では，気が不足した状態と過剰な状態を対象とし，不足と過剰のカテゴリは 5 段階に分類され，ランダムフォレストを用いて患者が該当するカテゴリを判別する．患者への質問は yes-no の 2 択で答えるものと，Visual Analogue Scale (VAS)の 2 種類がある．VAS は，患者が身体の部位ごとに感じている痛みを [0,100]で示すものであり，熱や痛みなどの物理的な刺激に対する個人の反応を把握することができ，治療に伴う患者の健康状態の変化を評価する．2830 人の患者の初診データを用いて学習を行った結果，テストデータに対しての識別率は 67.0%，20 項目以上に解答した患者のデータのみを利用の場合の識別率は 72.4%となった．また，証間の関係モデル化手法[4]では，証とその原因との関係を分析し，医師に提供することで患者に対する最適な治療法を決定する．しかし，この研究では，患者に該当する証を完全に特定できていることが前提であるという問題が存在する．

2.2 課題と提案

2.1 の関連研究から，患者に該当する証を特定することが診断を支援するうえで必要であり，質問内容を設定することで患者が該当するカテゴリの判定を行ったが，証には様々な種類があるだけでなく，診断に用いられる問診票は患者自身の言葉で記入する項目が存在するため，関連研究を用いて患者の証を完全に特定することは困難である．したがって証を特定するためには患者が問診票に記入した表現の異なる同義な文章を考慮する必要がある．また，証の名称や症状を表現する多数の中医学の専門用語について類似度の算出を行うには，一般的な単語だけでなく中医学の専門用語にも対応した分散表現の獲得が必要となる．そこで，Word2Vec[5]で異なるコーパスを用いた分散表現の獲得や，獲得した分散表現に対してファインチューニングによる分散表現の調整を行うことで，患者や文献による表現の多様性や専門用語に対応した中医学の診断に適した分散表現獲得を検討する．これにより，患者が問診票に記入した文章に対して類似度算出を可能とする．

3. 中医学の診断を支援する分散表現

3.1 コーパスの異なるモデル

Word2Vec を用いた分散表現の獲得時のコーパスによる性能を確認するため，異なるコーパスを用いた Word2Vec モデルの作成を行った．Word2Vec では，固有のベクトルを持たなかった単語は未知語として同じベクトルを持つ単語として学習される．作成したモデルは次の通りである．

Wikipedia モデル: Wikipedia[6]に記載されている日本語記事の本文をコーパスとしたモデルである．単語が固有のベクトルを持つまでの単語の出現回数(以降，出現回数)は 20 回

表 1. モデルが持つ単語の種類

モデルの名称	出現回数	一般的な単語	中医学単語
Wikipediaモデル	20	○	×
	5	○	×
中医学モデル	20	×	×
	5	△	○
複合モデル	5	○	○

表 2. 性能比較実験結果

	旧複合モデル		新複合モデル	
	症状	類似度	症状	類似度
1	眠りが浅い	1.0	眠りが浅い	1.0
2	咽が乾く	0.7524	咽が乾く	0.7545
3	口が乾くが水分は欲しくない	0.7088	口が乾くが水分は欲しくない	0.7030
4	喉の渇きが強い	0.7029	喉の渇きが強い	0.6900
5	腹がシクシク痛む	0.6944	痰に血が混じる	0.6825
6	皮膚がどす黒い	0.6944	舌尖が紅い	0.6810
7	夜間によく目が覚める	0.6902	呼吸が荒い	0.6802
8	舌尖が紅い	0.6895	腹がシクシク痛む	0.6765
9	爪がもろい	0.6863	寒冷時に四肢末梢の冷えを自覚	0.6763
10	呼吸が荒い	0.6840	夜間によく目が覚める	0.6723

表 3. 性能評価実験結果

	お腹が痛い		腰が痛い		手足のしびれ	
	症状	類似度	症状	類似度	症状	類似度
1	腹がシクシク痛む	0.8438	腰がだるく痛む	0.9702	手足のほてり	0.9559
2	腰がだるく痛む	0.8208	腰がだるい	0.9508	四肢のしびれ	0.9218
3	胸が熱苦しい	0.8036	胸が熱苦しい	0.8526	四肢の痙攣	0.8493
4	咽が乾く	0.8034	腹がシクシク痛む	0.8492	下肢のむくみ	0.8377
5	消化が悪く腹にもたれる	0.7789	腰や膝がだるく無力	0.8332	四肢の冷え	0.8307
6	腰がだるい	0.7781	腰やひざがだるく無力	0.8134	節々の痛み	0.8226
7	口が乾くが水分は欲しくない	0.7760	首筋や肩背のこり痛み	0.8075	下半身の冷え	0.8096
8	顔色が悪い	0.7737	腰や足がだるく無力	0.8058	下肢の軽度のむくみ	0.8072
9	痰がからむ	0.7638	一般に腰から下肢が冷えることが多い	0.7798	首筋や肩背のこり痛み	0.8056
10	すぐに目が覚めておびえる	0.7502	腰や四肢がだるく無力	0.7777	まぶたや顔面の筋肉がピクピクとひきつる	0.8041

と 5 回として学習を行った．このモデルでは，一般的な単語は固有のベクトルを持つが，「腎虚」などをはじめとした中医学の単語が未知語として判断された．

中医学モデル:中医学文献 2 冊[7][8]のテキストデータをコーパスとしたモデルである．このモデルは，出現回数 20 回のとき，コーパス内のほとんどの単語が未知語として処理されたため，出現回数を 5 回として学習を行った．その結果，一般的な単語に未知語として判断されるものが含まれたが，中医学の証や症状などの単語は固有のベクトルを持つことが確認できた．

複合モデル: Wikipedia に記載されている日本語記事と中医学文献をコーパスとしたモデルである．中医学モデルの学習結果から，中医学の単語が出現回数 5 回で固有のベクトルを持つことが確認できたため出現回数 5 回で学習を行った．その結果，一般的な単語だけでなく中医学の単語も固有のベクトルを持つことが確認できた．

また，それぞれのモデルで固有のベクトルを持つ単語の種類を表 1 に示す．表 1 より，Wikipedia モデルは一般的な単語，中医学モデルは中医学単語に特化したモデルであることが確認できる．これに対し，複合モデルでは，一般的な単語だけでなく，中医学の単語についても類似度を測定することができることを確認した．

3.2 コーパスの改良による性能変化

3.1 で作成した複合モデルのコーパスに中医学文献[9]を追加し，モデルの性能に与える変化の確認を行った．作成した新しい複合モデル(以降，新複合モデル)と 3.1 の複合モデル(以降，旧複合モデル)を用いて性能比較実験を行った．患者の症状を「眠りが浅い」とし，中医学文献[10]に示される証の症状との類似度を算出することでモデルの性能の差を確認する．性能比較実験の結果の上位 10 症状を表 2 に示す．旧複合モデルと新複合モデルの両方で「夜間によく目が覚める」という同じ内容を指す症状が上位の証に含まれていることを確認した．しかし，類似度上位には，患者の症状と類似していない症状が多く含まれている．また，新複合モデルは旧複合モデルよりも固有のベクトルを持つ単語の数が約 1,000 語増加している．このことからコーパスに文献を追加することで未知語を減らすことはでき，多くの単語に対しての比較が可能となる．

3.3 性能評価実験

3.2 で作成した新複合モデルの評価を行うため，患者の症状を仮定し，類似度算出実験を行った．類似度算出実験では，患者の症状を「お腹が痛い，腰が痛い，手足のしびれ」とし，中医学文献[10]に示される証の症状との類似度を算出し，類似度上位の症状からモデルの評価を行う．類似度算出実験の結果で類似度上位となった症状を表 3 に示す．患者の症状として想定した 3 つの症状の全てにおいて，最も類似度が高い症状が患者の症状と表現が異なる近い症状であるなど，患者の症状と類似した症状が類似度上位に算出されていることが確認できた．しかし，「お腹が痛い」を入力症状としたときに胸や腰の症状が上位に来るなど，患者の症状に該当しない部位の症状が上位に算出されたことが確認できた．この実験から，新複合モデルを用いて類似度算出を行うことで表現の多様性を考慮した症状の抽出の可能性を確認した．しかし，現状のモデルでは，症状が起こる部位の違いを考慮した類似度算出ができないなどの問題が存在する．コーパスの増加による改良では 3.2 よりモデルが持つ単語の数は増加するが，部位が考慮できないといった類似度の算出の傾向は変化しないことが分かっている．そこで，類似度算出の性能を改良する方法をして，モデル内で身体の部位を表す単語の距離を離すことを検討している．具体的には，部位を表す単語は文献や患者の表現の文章の中での使われ方が似ていることが原因で距離が近く，類似度が高い状態にある．そのため，身体の部位を表す単語のモデル内での距離を離すことで部位を含む文章でより正確な類似度が算出できると考える．

4. おわりに

本研究では，表現の多様性や専門用語を考慮した分散表現の獲得を目的とし，中医学文献を用いた分散表現獲得手法を検討している．本稿では，中医学文献を利用した分散表現の獲得と性能の確認を行い，表現の多様性と専門用語を考慮した症状抽出の可能性を確認した．今後は，獲得した分散表現に対し，単語間の距離を離すなどの調整を行うことでより本手法に適した分散表現を獲得する．

【参考文献・参照】

[1] ICD-11 for Mortality and Morbidity Statistics；https://icd. who. int/browse11/l-m/en(参照 2021,06,01)
[2]『中医学伝統医学 問診表』，東北大学病院漢方 内科，関隆志，2010
[3] Katayama, Kotoe & Yamaguchi, Rui & Imoto, Seiya & Watanabe, Kenji & Miyano, Satoru. Analysis of QuestionnaireforTraditional Medicine and Development of Decision Support System. Evidence-based complementary and alternative medicine (2014).
[4] 中国伝統医学（中医学）情報共有支援のための証間の関係モデル化手法, 五十嵐文, 情報処理学会論文誌 Vol.61 No.3 667- 675 (Mar. 2020)
[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. andDean, J.”Distributed Representations of Words and Phrases and their Compositionality”2013
[6] Wikipedia: https://ja.wikipedia.org/wiki
[7] 呉澤森, 高橋楊子,『「証」の診方・治し方』,2013
[8] 金子朝彦,『問診のすすめ』，2014
[9]『中医基本用語辞典』，中医基本用語辞典翻訳委員会(翻訳), 高金亮，劉桂平，孟静岩,2006
[10]『漢方・中医学臨床マニュアル』，森雄材,2004