6R-05

伝統医学分野における Sentence-BERT を用いた文献検索システムの検索性能の評価

伊藤 元斗† 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

2019 年に WHO が公表した国際疾病分類第 11 版に、伝統医学に関する章が追加された。高齢化が進む中、伝統医学特有の疾病を未然に防ぐ概念や、治療法に関する知見を活用し、健康寿命の延伸等に役立てることが期待されている。

伝統医学では、「証」と呼ばれる概念を取り扱う。証は患者の心身の状態を表し、医師の診察によって特定される。その過程では、複数の症状と証同士の因果関係を適切に結び付ける必要がある。これには多くの診療経験と知識が必要であり、診療経験の少ない医師を支援する仕組みが求められている。

そこで、証に基づく診断の支援を目的として、医師の知識を補完するための証の検索システム、システムで用いる証情報の DB が構築された。

2. 関連研究

証の検索システム[1]は、伝統中医学の 75 の証について Word2Vec を用いたベクトル検索を行うシステムである。Word2Vec は、単語の分散表現を得る機械学習ベースの手法であり、単語分散表現は文字列である単語を N 次元の実数のベクトルで表現したもので、意味や特徴が近い単語同士は似たベクトルとして表現される。各証の説明文の各単語の分散表現の平均値と、検索文字列の単語分散表現の平均値のコサイン類似度を計算し、類似度が高い順に証情報を並び替えて検索結果を提供する。伝統医学文献に記載される用語は、文献ごとの表記ゆれや一般的な用語が説明に用いられないこと等があり、一般的な語彙を検索文字列とする文字列一致ベース分散表現の獲得や検索手法では、適切な対象を検索結果として提供できない場合がある。そこで、機械学習ベースの分散表現を用い、検索時の表記ゆれに関する課題を解消した。表記ゆれのある検索対象については高い検索性能を発揮したが、それ以外の項目については TF-IDF に及ばず、より品質の高い文脈を考慮した分散表現が必要だとされた。

文脈を考慮した文の分散表現を得る手法として、Sentence-BERT[2](以降 SBERT と略記)が挙げられる。SBERT は BERT[3]と呼ばれる手法を文の分散表現を得る目的に特化させた手法である。文の類似度を判定するタスクにおいて単語分散表現を用いる手法よりも性能が高く、Word2Vec よりも高い検索性能を得られる分散表現の獲得が可能であると考えられる。

Evaluation of a Sentence-BERT based Literature Search System in Traditional Medicine
†National Institute of Technology, Sendai College
‡Fuji-Toranomon Orthopedic Hospital

3. 研究目的

本研究は、検索性能が高く、表記ゆれの問題を解消した、既知の伝統医学文献に対する検索システムを構築することが目的である。本稿では、SBERT を用いた伝統医学文献の文の分散表現の獲得と、証情報の検索性能の評価について述べる。

4. 研究手法

SBERT のモデルは、事前学習済みの BERT のモデルに対し、複数の文を入力としたファインチューニングを実施し得ることができる。本研究では NLI データセットを用いた学習手法を採用する。

4.1. NLI(自然言語推論)データセットの拡張

NLI データセットは、文のペア(前提文・仮説文)とその関係を表すラベル(entailment・neutral・contradiction)が一組となったデータで構成されている。本研究では、JSNLI データセット[4]を用いる。

また、伝統医学の情報を学習させるため、表 1 の構造を持つ証情報 DB を使用し、表 2 に示す形式でデータの拡張を実施した。entailment ラベルのデータ拡張は含意関係があると考えられるペアを選択した。neutral ラベルのデータ拡張は、いずれの症状も同時に発生し得る別々の事象という考えに基づいている。

表 1. 証情報 DB のデータ構造

項目名	内容	データ型
name	証の名前	文字列
description	証についての説明	文字列
cause	証の原因の説明	文字列
symptom	証の症状の説明	文字列
symptoms	証の症状のリスト	文字列のリスト

表 2. データ拡張の方法と例 (表証の場合)

entailment (4608 件)	ペアの生成例
name,symptom のペア	“表証の症状”, “悪寒、または…”
name,symptoms のペア	“表証の症状”, “浮脈” “表証の症状”, “咳” …
neutral (30478 件)	ペアの生成例
symptoms 内の2つの症状の組み合わせ	“浮脈”, “咳” “咳”, “頭痛” …

4.2. Sentence-BERT の学習

Multiple Negatives Ranking Loss による学習を行った。NLI の含意関係の認識結果を直接学習するのではなく、NLI データセットから作成した 3 つの文(A, B, C)の組を入力として、A と B の文の分散表現を近づ

け、C を遠ざけるような学習を実施する。遠ざける文 C の対象には、contradiction のラベルの文に加え、neutral のラベルの文を選択した。これは、証情報 DB を用いて拡張したデータを学習させるための操作であり、STS タスク等の性能は低下する可能性がある。

事前学習済みの BERT のモデルは、cl-tohoku/bert-base-japanese-v3[5]に対して未知語の削減を目的として 257 個の語彙の追加を行い、約 3.7MB の伝統医学文献由来のコーパスで継続して事前学習を行ったモデルを使用した。既知の文書に対する性能の良い情報検索の実現のために、コーパスには証情報 DB から得た文(約 200KB)を含めている。

BERT に対しては、学習率 5e-5、バッチサイズ 8、エポック数 200 で継続して事前学習を実施した。SBERT の学習時のパラメータは、学習率 2e-5、バッチサイズ 64、エポック数 1 とした。

5. 検索性能の評価

TF-IDF、Word2Vec[1]、SBERT の検索性能を比較するため、先行研究で伝統医学の専門医により作成された証情報検索のデータセット[1]を用いて、手法ごとに Recall@k と MAP を算出する。

5.1. 評価に用いるデータセット

75 の証について、5 つの症状(“いらいら”,“不眠”,“出血”,“浮腫”,“発熱”)ごとの正解/不正解が記載されており、不正解が多い不均衡なデータセットである。

5.2. 検索アルゴリズム

75 の証の症状の説明文と、5 つの症状から分散表現を獲得し、症状の分散表現と、75 の証の症状の説明文の分散表現のコサイン類似度を計算し、類似度が高い順に並び替えたもの検索結果としている。

5.3. Recall@k の計算

全体の正解の中で、検索結果の上位 k 件の中に含まれる正解の数を示す指標である。本稿では k=5,10 の場合を取り扱い、ユーザーが最初に閲覧する検索結果に適切な結果が含まれているかを評価する。式(1)により計算する。

$$Recall@k = \frac{\text{上位 k 件に含まれる正解の数}}{\text{全体の正解の数}} \quad (1)$$

5.4. MAP(Mean Average Precision)の計算

検索結果全体の精度を評価する指標であり、症状に関連する証が適切なランク付けをされているかを評価する。MAP は式(2)、AP(Average Precision)は式(3)により計算する。

$|Q|$ = 検索文字列の数 (今回は 5)

n = 検索対象のドキュメントの数 (今回は 75)

$rel(k)$ = 上から k 番目の結果が正解なら 1、不正解なら 0 を返す関数

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AP_i \quad (2)$$

$$AP_i = \left(\sum_{k=1}^{n} \frac{\text{上位 k 件に含まれる正解数}}{k} \times rel(k)\right) \div \text{全体の正解の数} \quad (3)$$

5.5. 評価結果

表 3,4 より、TF-IDF の場合、出血に対応する検索結果が上位 10 件以内に含まれていないことが確認できる。出血という文字列は症状の説明文に直接含まれておらず、機械学習ベースの手法により分散表現を得る場合は、出血に対応する証情報を得られていることから、Word2Vec や SBERT は表記ゆれの問題を解消していると考えられる。また、表 3,4,5 より SBERT は全体的に TF-IDF と同等、もしくはそれ以上のスコアを示しており、既知の文書の検索に対して、高い検索性能を有していると考えられる。

表 3. 各手法、各症状の Recall@5

手法	いらいら	不眠	出血	浮腫	発熱
TF-IDF	0.25	0.25	0.0	0.5	0.17
Word2Vec	0.25	0.19	0.5	0.2	0.17
SBERT	0.31	0.31	0.5	0.5	0.17

表 4. 各手法、各症状の Recall@10

手法	いらいら	不眠	出血	浮腫	発熱
TF-IDF	0.56	0.44	0.0	0.7	0.38
Word2Vec	0.31	0.25	0.5	0.3	0.29
SBERT	0.63	0.63	1.0	0.7	0.38

表 5. 各手法の MAP と、各手法、各症状の AP

手法	いらいら	不眠	出血	浮腫	発熱	MAP
TF-IDF	0.75	0.59	0.04	0.82	0.80	0.60
Word2Vec	0.63	0.39	0.52	0.36	0.68	0.52
SBERT	0.94	0.91	0.67	0.83	0.87	0.84

6. まとめ

本研究では、既知の伝統医学文献に対する、高い性能を持つ検索システムを構築することを目的に、SBERT のモデルを作成し、それを用いた検索性能の評価と先行研究[1]との比較を行った。

SBERT による 75 の証情報についての検索は、全体的に TF-IDF と同等かそれ以上の性能を発揮し、表記ゆれのある対象についても検索結果に含めることが可能である。

謝辞

本研究は JSPS 科学研究費補助金(JP23K11344)の助成を受けたものです。

参考文献

[1]太田遥人, 関隆志, 高橋晶子, 力武克彰:中医学のための単語埋め込みに基づく情報検索システムの研究, 情報処理学会 第 84 回全国大会講演論文集 2022(1), pp.747-748, (2022)

[2]Nils Reimers, Iryna Gurevych : Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp.3982-3992, (2019)

[3]Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp.4171-4186, (2019)

[4]吉越 卓見, 河原 大輔, 黒橋 禎夫:機械翻訳を用いた自然言語推論データセットの多言語化, 第 244 回自然言語処理研究会, 情報処理学会研究報告, Vol. 2020-NL-244, No.6, pp.1-8, (2020)

[5]東北大学自然言語処理研究グループ, "cl-tohoku/bert-base-japanese-v3 - Hugging Face", URL:https://huggingface.co/cl-tohoku/bert-base-japanese-v3