参考文献分類タスクのデータセット整備とクラスタリングを用いた分類手法の検討

細川諒真[1] 大和淳司[1] 東中竜一郎[2]
[1]工学院大学 情報学部
[2]日本電信電話株式会社
j318244@ns.kogakuin.ac.jp, jy@acm.org
ryuichiro.higashinaka.tp@hco.ntt.co.jp

概要

科学論文執筆支援を目的として，関連研究をまとまりに分ける手法に取り組んでいる． 本研究では，論文の参考文献とパラグラフが対応したデータセットを作成し BERT, BioBERT, SciBERT の各言語モデルを利用して参考文献のクラスタリングを行った．論文のデータセットはオンライン論文アーカイブである PMC から取得した論文を使用して作成した．各 BERT モデルで参考文献の Embedding を行いクラスタリングを行った場合のクラスタリング精度は BioBERT が BERT よりも優位であったが，2,200 論文分の参考文献のペアのクラスタ異同を学習させると BERT による手法の精度は向上し，モデル間のクラスタリング精度差は減少することを示した．

1 はじめに

近年，科学論文の数は急激に増加しており，科学論文執筆の補助を行うことについての研究も盛んに行われている．成松らの論文では，科学論文執筆支援におけるタスクを定義しており，それぞれの段階で必要なタスクについてまとめられている[1]．本研究では，成松らの定義した引用文献分類タスクに取り組む．このタスクは， 論文で引用している文献リストを対象に，これらをカテゴリ別に分類するタスクである．これは，論文執筆支援において，論文の参考文献群が与えられたときに，それぞれの文献を適切なクラスタに分類し，関連研究のセクションを書きやすくすることを目的とするタスクである．本稿では， このタスクの精度向上のために，参考文献のクラスタリングを行うためのデータセットの作成と，データセットを用いてクラスタリング実験を行う．

2 提案手法

本研究では，実際の論文データを収集し，その論文の関連研究のパラグラフのそれぞれで引用されている参考文献を列挙することで，参考文献のクラスタの正解データを作成する．また，参考文献群の論文番号と Abstract をベクトル化したデータとの対応したデータセットを作成する．参考文献群の Abstract のデータをもとにクラスタリングを行い，実際の論文のパラグラフと生成したクラスタとの精度を評価する．本研究では自然言語処理モデルとして BERT とそれを利用した SciBERT, BioBERT の 3 種類を使用し，文書のベクトル化を行う BERT は，2018 年に開発された大規模なデータを用いる双方向 Transformer による事前学習モデルである[2]．SciBERT は，Semantic Scholar の科学論文を 114 万件用いて学習を行った BERT モデルである[3]．BioBERT は，オリジナルの BERT と同じ英語版の Wikipedia, BookCorpus, に加えて PubMed に掲載されている論文のアブストラクトから取得した 45 億単語と PMC の論文から取得した 135 億単語を用いて学習を行ったモデルである[4]．本研究では，PubMed の論文で 100 万ステップ学習した BioBERTv1.1 を利用する．学習に利用しているデータが SciBERT とは異なり，BioBERT は，より生物医学領域に特化している．

3 実験手順

3.1 データセット作成

本研究では，論文のデータセットとして医学分野のオンライン論文アーカイブである PMC から論文

データを取得した. 対象とする論文は, 著者らが進めている関連テーマの研究領域をモデルケースとした. 具体的には, PMC から Artificial Intelligence , Deep Learning, Image Recognition に関する論文の xml データを 194,787 件取得した. 取得したデータの中から article-type が research-article であり, Background のセクションで始まる論文を 12,178 件抽出した. PMC の論文では, article-type が case-report となっている症例報告が多数あるためこのような絞り込みを行った. また, 抽出した論文に参考文献として登録されている論文の Abstract を PubMed から 122,555 件取得した. 12,178 件の論文データに対して, コンピュータ科学分野における関連研究のセクションに対応していると考えた Background のパラグラフごとに対応した参考文献群のデータを作成した. 作成したパラグラフと参考文献群の対応データの内, 参考文献の登録されているパラグラフが 2 件以上であり, 参考文献の Abstract が PubMed から取得できていない参考文献が 2 件以下であり, 取得できない参考文献によりパラグラフに対応した参考文献が 0 件になることのないデータを 2,752 件作成した. 参考文献数の平均は 15 件であり, PubMed から取得できていない参考文献が 2 件までならば妥当なクラスタリングができると判断した.

また, 参考文献群の 122,555 件の Abstract に対して BERT, SciBERT, BioBERT を使用して文書の Embedding を行いそれぞれ 768 次元のベクトルに変換したデータセットを作成した.

表 1 参考文献群データの例

元論文のPMCID	PMC6371455
パラグラフ1	[1, 2, 3, 4, 5, 6, 1]
パラグラフ2	[7, 5, 8, 9, 10, 11, 9, 10, 11]
パラグラフ3	[8, 12, 13, 14]
文献1のPMID	12626338
文献2のPMID	23258890
文献3のPMID	22909801
文献4のPMID	20573213
文献5のPMID	25525159
文献6のPMID	18688268
文献7のPMID	27717327
文献8のPMID	28655331
文献9のPMID	21964334
文献10のPMID	23938295
文献11のPMID	25704815
文献12のPMID	28673540
文献13のPMID	26740580
文献14のPMID	26752769

表 1 に作成したパラグラフと対応した参考文献群データの例を示す.

3.2 手法

本研究では, 実験を 3 つ実施した.一つ目は, 参考文献の Abstract のベクトルと用いた手法である. 二つ目は, BertForSequenceClassification を利用して BERT モデルに参考文献 2 つが同じクラスタに所属しているか学習させる方法である. 三つ目は, 参考文献の Abstract ベクトル 2 つに元論文の Abstract ベクトルを連結し多層パーセプトロンに学習させる方法である.

3.2.1 Abstract のベクトルを用いた手法

実験 1 では, 3.1 節で作成した論文データセットに対して, クラスタ数を与えて k-means 法を利用して参考文献の Abstract のベクトルを用いてクラスタリングを行う. クラスタリング精度の評価として, 佐藤らの論文[5]でも用いられていた Purity を用いる. Purity の定義式は, N をクラスタリングの要素数, $\{C_i\}$は生成した各クラスタ, $\{A_j\}$は正解のクラスタとすると以下の式で定義される.

$$Purity = \sum_i \frac{|C_i|}{N} \max_j Precision(C_i, A_j)$$

$$Precision(C, A) = \frac{|C \cap A|}{|C|}$$

本実験では, Abstract のベクトルとして 3.1 節で作成した BERT , BioBERT , SciBERT の 3 種類のベクトルをそれぞれ利用してクラスタリングを行う. 比較用に参考文献のクラスタを要素 0 のクラスタが発生しないようにランダムに割り振った Random のクラスタリングを行う. 実験 1 では, データセット全てに対する 2,752 論文のクラスタリングと実験 2, 3 と同じテストデータの 276 論文のクラスタリングを行う.

3.2.2 参考文献間の距離を用いた手法

実験 2 では, BertForSequenceClassification を利用して, 3 種類の BERT モデルに論文の参考文献 2 つの Abstract とその 2 つの参考文献が同じクラスタに所属しているかどうかを表すラベルとの関係を学習させた. このラベルは 1 ならば同じクラスタに所属していることを表し, 0 ならば同じクラスタに所属していないことを表す. 本実験では, 2,752 件の論文の内 2200 論文から作成した参考文献のペア 360,834 件を学習データ, 276 論文から作成した参考文献のペア 46,570 件をテストデータ, 276 論文から作成した 45,754 件を検証データとした. 学習データの内 38%になる 138,083 件がラベル 1 であり, 残りの

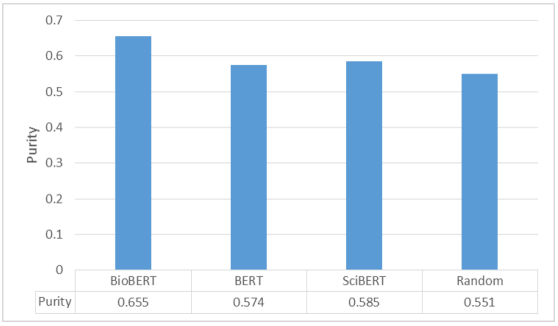

図 1 実験 1 の 2752 論文に対する クラスタリング精度の評価

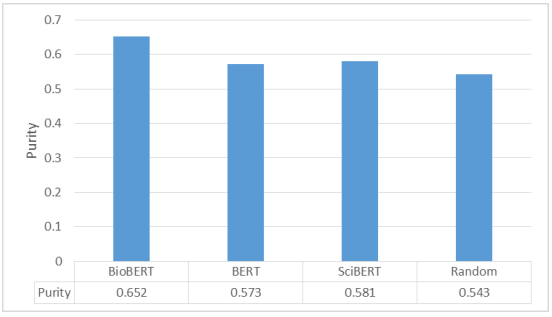

図 2 実験 1 の 276 論文に対する クラスタリング精度の評価

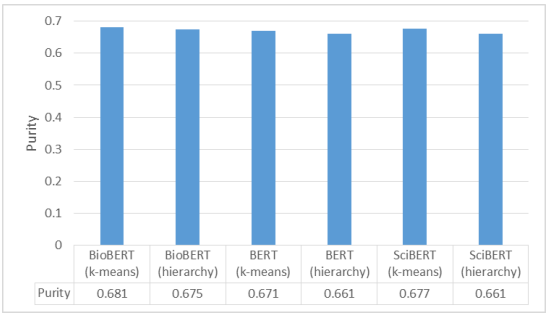

図 3 実験 2 のクラスタリング精度の評価

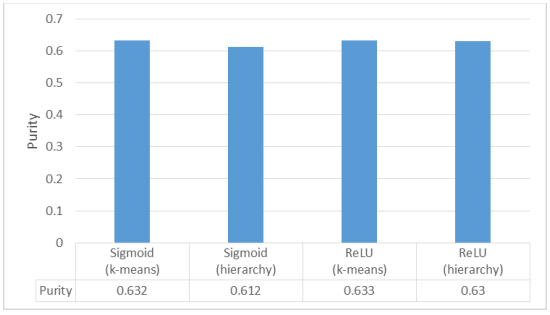

図 4 実験 3 のクラスタリング精度の評価

62%がラベル 0 であり若干の偏りが見られるため, クラスの重みを付け学習させた. 学習後の BERT モデルに対して, テストデータの論文の参考文献を 2 つ入力することでラベルの予測に使用する logit0, logit1 の値を取得する. logit0 と logit1 の値の高いほうがラベルの値の予測結果となる. logit0 と logit1 の値にソフトマックス関数を適用することで正規化した logit0 の値を参考文献間の距離とした. この参考文献間の距離値を用いて, k-means 法と階層的クラスタリングを行う. Purity を用いてクラスタリング精度の評価をする.

3.2.3 多層パーセプトロンを用いた手法

実験 3 では, 3.1 節で作成した BERT による参考文献 2 つの Abstract と元論文の Abstract を新たに BERT で変換して得たベクトルを連結し, 2,304 次元のデータとして 2 つの参考文献が同じクラスタに所属しているかどうかを表すラベルとの関係を学習させる. 本実験では, 実験 2 と同じ 2,200 論文から作成した 360,834 件を学習データとし, 実験 1, 2 と同じ 276 論文をテストデータとする. 入力層に 2,304 ノード, 中間層に 1,000 ノード, 出力層に 2 ノード配置したネットワークを構成する. 損失関数には Negative Log Likelihood (NLL) Loss を使用する. 中間層の活性化関数にシグモイド関数と ReLU 関数を用いた 2 つのモデルを作成し, 出力はログソフトマックス関数を使用した. 参考文献間の距離を用いて, k-means 法と階層的クラスタリングを行う. Purity を用いてクラスタリング精度の評価をする.

4 実験結果

表 2 実験 2 のクラスタリング例 1

正解	予測
[1, 2, 3, 4, 5]	[3, 4, 17]
[5, 6, 7, 8, 9, 10]	[1, 2, 5, 6, 7, 8, 9, 10]
[11, 12, 13, 14, 15, 16,	[11, 12, 13, 14, 15, 16]

Purity=0.824

表 3 実験 2 のクラスタリング例 2

正解	予測
[1, 2, 3, 4, 5, 6]	[1, 2, 3, 4, 5, 7]
[5, 7, 8, 9, 10, 11]	[6, 9, 10, 11, 14]
[8, 12, 13, 14]	[8, 12, 13]

Purity=0.786

図 1, 2, 3, 4 に各実験のクラスタリング精度を Purity で評価したときの結果を示す. 図 3, 4 中の hierarchy は, 階層的クラスタリングによってクラスタリングした際の結果を Purity で評価した結果を示す. 表 2, 3 に, クラスタリング例を示す. Purity と実際の正解のクラスタに対応した, 予測のクラスタを右に示す. 予測クラスタのオレンジ色は正解クラスタと同一のものを表す.

5 考察

図 1, 2 の結果より, 3.2.1 節の参考文献の Abstract を BERT モデルごとにベクトルへ変換したクラスタリングの精度は, BioBERT を利用してベクトルに変換したものが最も高いことが確認できる. これは, 本実験で用いた論文データが PMC の医療論文であるため, 医療分野の語彙に特化したモデルである BioBERT の方が他の 2 つのモデルより文章の特徴をつかむことができたからだと考えられる. 図 3 より, 実験 1 より高い Purity を示したことから, 学習することによりクラスタリングの精度が上昇したことが確認できる. 実験 1 では各 BERT モデル間の Purity の差が大きかったが, 実験 2 では BERT モデル毎の差が小さくなっていることが確認できる. このことから, 実験 2 で 2,200 論文分の 360,834 件の文章のペアを学習することでモデル間の語彙の差が縮まり, ラベルの予測の差が付きにくくなったと考えられる. 図 4 の結果より, 学習データに論文の Abstract を追加して 2,304 次元のデータとして学習させたが, 実験 2 の各モデルより高い精度ではないことを示した. このことから, 実験 3 では適切なモデルを構成できていない可能性があることが考えられる.

また, 本研究のクラスタリングでは, 1 つの参考文献は 1 つのクラスタにしか所属することはないが, 実際の論文では, 参考文献が複数のクラスタに所属することもある. 複数クラスタに所属している要素が存在した場合, その要素のクラスタを予測する際には該当するどのクラスタに分類しても正解となってしまう. そのために, Purity の値を高く見積もっている可能性があり, 今後の検証が必要である.

6 おわりに

本研究では, 参考文献のクラスタリングを行うためのデータセットを作成し, 3 つのクラスタリング手法を提案した. BERT モデルに参考文献 2 つと同じクラスタに所属しているかどうかのラベルを学習させることで, クラスタリング精度が上昇することを示した. 今後の課題としては, 参考文献の重複を許すクラスタリングを行うことや, 人間によるクラスタリング結果との精度の比較などが考えられる. クラスタリング精度の評価に関しては, 本研究では Purity のみを用いて評価してきたが, 他に, Inverse Purity と F 値が知られている. これらの値を用いて評価することも可能であると考えられる. BERT モデルに関しては, 本研究では 3 種類のモデルを使用した. 他のモデルを用いることでクラスタリング精度の上昇を図ることもできると考えられる.

謝辞

ご議論いただいた NTT コミュニケーション科学基礎研究所杉山弘晃氏, 成松宏美氏, 農研機構菊井玄一郎チーム長, 大阪工業大学平博順准教授, 電気通信大学南泰浩教授, 秋田県立大学堂坂浩二教授に感謝します.

参考文献

[1] Hiromi Narimatsu, Kohei Koyama, Kohji Dohsaka, Ryuichiro Higashinaka, Yasuhiro Minam, Hirotoshi Taira, “Task Definition and Integration for Scientific-Document Writing Support”, Proceedings of the Second Workshop on Scholarly Document Processing, pp.18-26, June2021.

[2] 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 (Long and Short Papers), pages 4171‑4186. May2019.

[3] Iz Beltagy, Kyle Lo, Arman Cohan, “SciBERT: A Pretrained Language Model for Scientific Text”, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615 ‑ 3620, Nov2019.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang,” BioBERT: a pre-trained biomedical language representation model for biomedical text mining”, arXiv:1901.08746v4, Oct2019.

[5] 佐藤 進也, 高橋 公海, 松尾 真人, “特徴抽出を目的とした文書クラスタからの一貫性阻害要素除去”, 情報処理学会論文誌データベース（TOD）6(3), 1-12, 2013-06-28.