Fusion-in-Decoder を用いた論文分類手法の検討

奥田あずみ　美野秀弥　後藤淳
NHK 放送技術研究所
{okuda.a-gc,mino.h-gq,goto.j-fw}@nhk.or.jp

概要

テキストを適切なカテゴリに分類するテキスト分類技術は言語処理において重要な技術の 1 つである．特に，近年，大量のテキストデータを扱うことが増え，テキスト分類の需要は高まっている．そこで，本稿では，大量の論文を分析するための論文のラベル分類タスクに取り組む．質問応答タスクで用いられている Fusion-in-Decoder の手法を参考に，論文などの長いテキストに対しても高精度のラベル分類できるような手法を提案した．論文のラベル分類タスクで評価実験を行った結果，提案手法の効果を確認した．

1　はじめに

テキストを適切なカテゴリに分類するテキスト分類技術は言語処理において重要な技術の 1 つであり，多くの手法が提案されている [1]．特に，近年は計算機などのハードウェアの技術進歩により大量のテキストデータが扱えるようになり，テキスト分類の需要は高まっている．最近では，新型コロナウイルスに関する論文が大量に公開され[1)]，それらの論文の迅速な解析が求められた．そこで，本研究では，論文のラベル分類タスクに取り組む．

BERT [2] などの事前学習モデルには入力長の制限値が設定されており，論文などの文長の長いデータを一度に扱えない課題がある．この課題への対応として，論文の全文を事前学習モデルの入力長の制限値以下に分割して複数のラベルを出力し，複数のラベルを結合して単一ラベルを出力する手法が提案されているが [3]，分割された各データのラベル分類に対する重要度を考慮しておらず，十分な精度が達成できていない可能性がある．そこで，本研究では，長文のテキストのラベル分類の高精度化のために，質問応答 (QA) タスクで用いられている手法を応用した手法を提案する．具体的には，論文の重要な情報が集約されていると考えられる概要部分を用いて論文の全文からラベル分類に有効な文集合を抽出し，抽出した文集合の各文のベクトル表現を結合してラベルを推定する．質問応答タスクで高い性能を達成している Fusion-in-Decoder [4] を用いることで，ニューラルネットワークを用いたベクトル表現の結合を実現した．新型コロナウイルスの科学技術論文データ CORD-19 [5] を用いた評価実験を行い，提案手法の効果を確認した．

2　関連研究

論文のラベル分類の手法に Sentence-Label-Classification(SLC) [3] がある．SLC は学術論文のような長い文章に対し，BERT [2] の入力長の制限値 (512 トークン) 以下になるように 1-3 文程度の短い文集合に分割し，その単位ごとに BERT モデルを用いて追加学習を行う．分割した文集合それぞれでモデルを学習するため学習データが増え，低頻度ラベルの学習に効果がある．しかし，この手法は，ラベルの統合時に各出力の平均値などを用いており，各文の重要度を考慮できていないため，改善の余地がある．

本稿では，SLC を改善するために，質問応答 (QA) タスクで用いられている手法を参考にして，ラベル統合時にニューラルネットワークを用いる手法を提案する．QA タスクの多くは，質問と回答を見つけ出すための文書集合から解答を見つけ出すタスクである．Karpukhin ら [6] は文書集合から関連する箇所を検索する Dense Passage Retriever (DPR) を提案し，オープンドメイン QA タスク[2)]で高精度な結果を達成した．Lee ら [7] は，オープンドメイン QA タスクに取り組み，Open-Retrieval Question

1）2022 年までに発表された新型コロナウイルス関連の論文数は，全文がオープンデータ化されているものだけでも 40 万本を超えている．

2）オープンドメイン QA タスクは学習データとして質問と解答のみが与えられ，推論時は Wikipedia のような外部知識を参照して解答の根拠を見つけ出すタスクである．

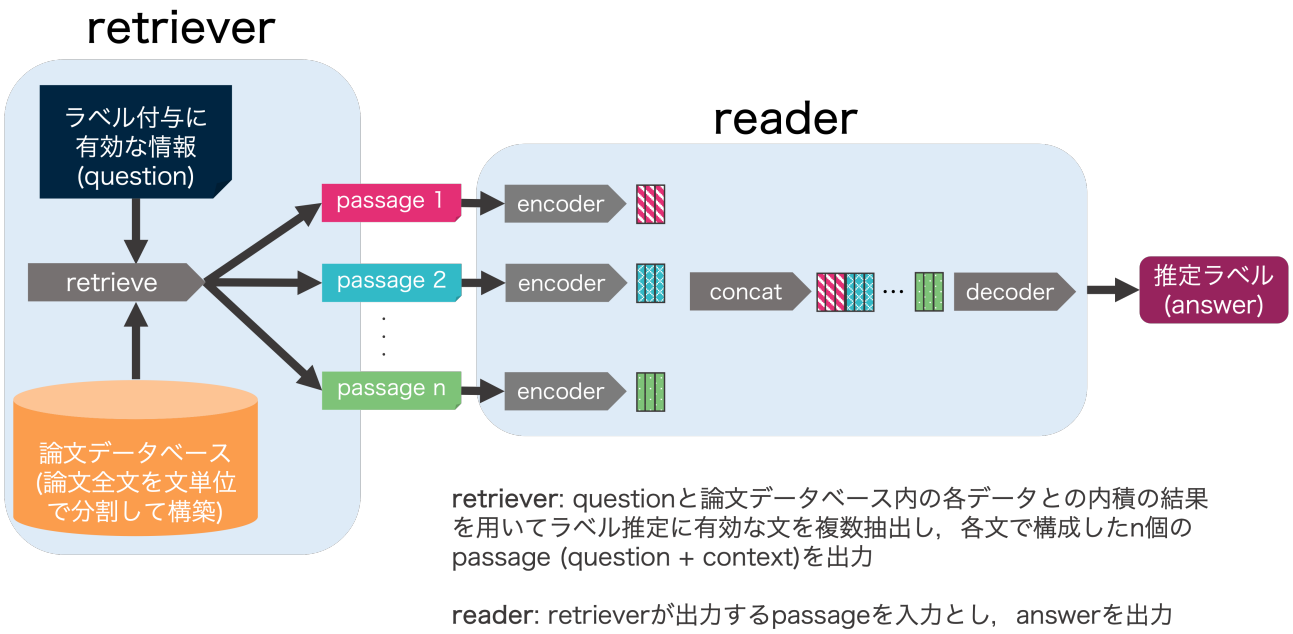

図 1　提案手法の概要

Answering (ORQA) モデルを提案した．ORQA モデルは，解答を得るのに適切な文集合を外部知識から検索する retriever と，retriever が検索した文集合と質問を合わせたものから解答を見つけ出す reader の 2 つで構成される．Izacard ら [4] は，ORQA モデルをベースに，reader 部分を改良して Fusion-in-Decoder (FiD) を提案した．さらに，Izacard ら [8] は reader の知識を用いて retriever を学習する手法を提案し，QA タスクの精度を向上させた．本稿では，Izcard らの FiD の手法を論文のラベル分類に応用している．

3　提案手法

3.1　概要

本研究では，Izcard らが提案した reader と retriever の 2 つの機構を持つ Fusion-in-Decoder (FiD) を，論文のラベル分類に応用する手法を提案する．提案手法の概要図を図 1 に示す．

reader は，ラベル分類に有効な情報 (question) と論文中のラベル分類に有効な文集合 (ctxs) を入力とし，論文のラベル (answer) を出力するモデルである．入力には，文集合中の各文 (context) に加え，context とは別の answer を推定するのに有効な情報である question を用いることができ，question と context を合わせたものを passage と呼ぶ．passage は複数用意することができ，それぞれ reader の encoder に入力して埋め込み表現を得る．そして，獲得した複数の出力結果を連結して decoder に入力して answer を生成するように学習する．

retriever は，論文全体を入力とし，ラベル分類に有効な文集合である ctxs を検索するモデルである．reader は複数の passage を encoder に入力してラベル分類を行う過程で，question と passage との cross-attention スコアが計算される．cross-attention スコアの高い passage は answer の推定に寄与しているという仮定のもと，retriever は，ラベル分類に有効な文集合を抽出するため，cross-attention スコアと question と passage の埋め込み表現の内積が関連付くように学習する．

reader が用いる passage は retriever により更新される．retriever は学習時に reader が出力過程で計算する cross-attention スコアの結果を用いる．reader と retriever は相互に関係しており，reader と retriever を繰り返し学習することでそれぞれのモデルを高精度化する．

3.2　手順

Izcard ら [4] は retriever の検索対象として，wikipedia のような大量のデータを含む外部知識を用いている．本研究では，図 1 にあるように，論文全文を文単位で分割して格納した論文データベースを外部知識とみなし，retriever を用いて論文データベース内の全文からラベル分類に有効な文を抽出する．

学習データは，短いテキストである論文の概要部分 (abstract)，長いテキストである論文全文 (fulltext)，論文のジャンルを表すラベル (answer)，の 3 つで構成される．

学習時の具体的な手順は以下の通りである．

1. **reader の学習**　学習データの abstract と fulltext を 1 文ごとに分割する．reader の encoder に分割した abstract を context として入力し，正解ラベルを answer として reader の最初の学習を行う．question には全論文共通で “What genre best describes this abstract?”[3] を用いる．
2. **retriever の学習**　訓練済みの reader が途中で計算する cross-attention スコアをもとに，question と context の内積が適切な値となるように retriever を学習する．
3. **retriever による推論**　2 で学習した retriever により論文データベースから関連する文を faiss [9] で最近傍探索を行い関連度が高い文を抽出する．極端に短い文は抽出しないようにした．この抽出作業を passage retrieval と呼ぶ．
4. **reader の再学習**　question と 3 で抽出した context とを reader に入力して再度 reader を学習する．
5. **交互に学習**　3-5 を繰り返し，reader と retriever を交互に学習する．

ラベル分類を行う際は，abstract を 1 文ごとに分割して最終的に得られた reader にかけ，出力結果の answer が推定ラベルとなる．

4　実験

4.1　実験設定

Allen Institute for AI が収集していた新型コロナウイルス感染症 (Covid-19) に関する科学論文リソースである The Covid-19 Open Research Dataset (CORD-19) を用いて実験を行った．CORD-19 は PubMed Central や WHO など，複数のリソースから学術文献検索サービス Semantic Scholar で検索して取得した論文を統合しており，著者名や投稿日時や取得ソースが abstract と併せてメタデータに記載されている．本稿では CORD-19 から抽出した bioRxiv に投稿された 7127 本の論文と，各論文に付与された 25 種類の研究分野のラベルを用いた．提案手法のモデルに入力するデータは付録 A のような JSON 形式のデータである．

context の集合である ctxs は retriever の結果を用いるが，初回の reader の学習時には retriever の学習は行われず用いることができない．そこで，ctxs の初期値として，論文の abstract がラベル分類に有効な文集合であると仮定し，学習データの各論文の abstract を文単位に分割したものを用いた．単語の分散表現には T5 (Text-to-Text Transfer Transformer) [10] を用いた．retriever の出力には論文データベースの検索結果から 10 単語以下で構成される文を除いた Top 20 を用いた．

比較手法には Sentence-Label-Classification(SLC) [3] を用いた．

評価には分類タスクで一般的に用いられている Micro-F1 と Macro-F1 を用いた．

表 1　実験結果．question には共通の文を用いた．

	context	繰り返し回数	Micro-F1	Macro-F1
SLC	-	-	0.480	0.370
提案手法	abstract	0	0.555	0.362
	retriever 結果	1	0.562	0.298
	retriever 結果	2	0.554	0.339
	retriever 結果	3	0.575	0.377
	retriever 結果	4	0.570	0.386
	retriever 結果	5	0.551	0.350

4.2　実験結果

実験結果を表 1 に示す．繰り返し回数は，reader と retriever の学習の回数を示す．

先行研究である SLC と繰り返し回数 0 回の提案手法を比較すると，Macro-F1 はほぼ変わらず，Micro-F1 は向上した．SLC は文ごとにラベル分類を行い，それらの結果をもとにルールベースでラベルを推定する．一方，提案手法は文ごとの encoder 結果を連結し，それを decoder で処理してラベルを推定しており，ニューラルネットワークを用いた手法である．提案手法が SLC より Macro-F1 で向上した要因として，ラベル分類にニューラルネットワークを用いる利点が考えられる．context 中のすべての文の重要度を考慮せずにラベル分類する SLC とは異なり，提案手法は文集合中のどの文がラベル分類に重要かを考慮したラベル分類が可能になっている．

次に，retriever と reader の学習を繰り返して，retriever による passage retrieval の効果を検証した．繰り返し 1 回目では，retrieve を用いない場合 (繰り返し回数:0) と比較して，Micro-F1 は若干向上し

3)　ラベル推定に必須の情報を question に入力することが好ましいと考えられるが，本研究では固定の表現を用いた．適切な question の検討は 5 章で分析する．

表 2　分析用の実験結果

	question	context	繰り返し回数	Micro-F1	Macro-F1
FiD	固定	abstract	0	0.555	0.362
FiD	固定	全文	0	0.590	0.380
FiD	abstract	abstract	0	0.570	0.396

たが，Micro-F1 は下がった．さらに繰り返し回数を重ねると，4 回目まで Micro-F1 はほぼ変わらず，Macro-F1 は向上した．5 回目で Micro-F1，Macro-F1 ともに数値が下がったため，学習の繰り返しを終了した．提案手法を用いた QA タスクにおける実験 [4] でも，4 回目程度まで性能が向上したと報告されており，本研究でも同様の傾向となった．

5　分析

本章では，retriever を用いない場合の reader の性能を分析する．

5.1　初期 context に全文を利用

提案手法では，reader で用いる context の初期値 (繰り返し回数:0) にラベル分類に有効な情報であると考えられる abstract を用いた．しかし，context には大量のデータを用いることが可能である．そこで，context の初期値に論文全文を入れて 4 章と同様の実験を行った．

表 2 の 2 行目が実験結果である．4 章の実験の繰り返し回数 0 回の結果 (表 2 の 1 行目) と比較して Micro-F1，Macro-F1 ともに精度が向上した．この結果より，context に入力するデータ数を増やすことで精度が向上すると考えられる．一方で，reader の context の初期値に論文全文を用いると計算に時間がかかるという課題がある．本研究の実験環境[4)]では，reader の学習に全文で約 9 時間，abstract のみでは約 3 時間かかった．

5.2　question に abstract を利用

4 章の実験の繰り返し回数 0 回の結果 (表 1 の 2 行目) は，retriever を用いずに abstract を用いてラベルを推定しており，abstract にラベル分類に有効な情報が含まれているかの判断材料となる．Micro-F1 の値は SLC (表 1 の 1 行目) よりも高いことから，Abstract にはラベル分類に有効な情報が含まれていると考えられる．そこで，question に共通の固定ベクトルを用いるのではなく，abstract の分散表現ベクトルを用いて実験を行った．表 2 の 3 行目が実験結果である．4 章の実験の繰り返し回数 0 回の結果 (表 2 の 1 行目) と比較して Micro-F1，Macro-F1 ともに精度が向上した．この結果より，question にラベル分類に有効な情報を入力することでラベル分類の精度が向上すると考えられる．

4) NVIDIA V100 1 コアを用いて実験を行った．

6　まとめ

本研究では，質問応答タスクで用いられている，retriever と reader からなる Fusion-in-Decoder を分類タスクに応用して，論文のラベル分類に取り組んだ．retriever では，文章全体の特徴を表す短い文と question を用いて長いテキストからラベル分類に有効な文集合を抽出し，reader では抽出された複数の文を統合し，ラベル分類を行った．新型コロナウイルス関連の論文を集めたデータセットを用いてラベル分類実験を行い提案手法の効果を確認した．今後は，より有効な question の選び方や単語埋め込み表現の改善を通じて，retriever の性能を上げる方法を検討する．また，異なるデータセットでの実験を行い，他の先行研究とも比較して提案手法の有効性を示したい．

謝辞

貴重なコメントや議論を頂いた NHK 放送技術研究所の山田一郎シニアリード，宮﨑 太郎研究員，安田有希研究員，石渡太智研究員に感謝します．

参考文献

[1] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: A comprehensive review. **ACM Comput. Surv.**, Vol. 54, No. 3, pp. 62:1–62:40, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] 奥田あずみ, 宮﨑太郎, 美野秀弥, 後藤淳. 文単位の分類器を用いた論文分類手法の検討. 2021 年映像情報メディア学会冬季大会, 2021.

[4] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 874–880, 2021.

[5] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[6] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics.

[7] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. In **The International Conference on Learning Representations**, 2021.

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. **IEEE Transactions on Big Data**, Vol. 7, No. 3, pp. 535–547, 2019.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

A 入力例


```
{
    "id": "047xpt2c",
    "question": "What genre best describes this
abstract?",
    "answers": [
        "pharmacology and toxicology"
    ],
    "ctxs": [
        {
                  "text": "A novel coronavirus
SARS-CoV-2, also called novel coronavirus 2019
(nCoV-19), started to circulate among humans
around December 2019, and it is now widespread
as a global pandemic.",
            "id": "047xpt2c0",
            "title": " "
        },
        {
                 "text": "The disease caused by
SARS-CoV-2 virus is called COVID-19, which is
highly contagious and has an overall mortality
rate of 6.96% as of May 4, 2020.",
            "id": "047xpt2c1",
            "title": " "
        },
        {
                "text": "There is no vaccine or
antiviral available for SARS-CoV-2.",
            "id": "047xpt2c2",
            "title": " "
        },
    ]
}
```