文書外の書誌情報と用語情報を組み込んだ文書分類

井田 龍希　三輪 誠　佐々木 裕
豊田工業大学
{sd22401,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

本論文では，書誌情報に基づく文献グラフを組み込んだ新しい文書分類手法を提案する．近年，事前学習モデルである BERT の利用により文書分類の性能は大幅に向上した．さらなる性能向上のためには，書誌情報や引用情報，用語の説明文や上位下位関係などの外部情報など，対象テキスト情報以外の情報の活用が鍵となると考えられる．本提案手法では，対象テキスト情報以外の情報を利用する既存の文書分類手法が考慮できていない，書誌情報や用語などの文書間に共通する情報を含む文書間の関係を文献グラフとして利用し，文書間のより詳細な関係を考慮する．実験では，生化学分野の文書分類データセットで既存手法からの性能向上を確認した．

1　はじめに

近年提案された，大量のデータで事前学習をし，文脈を考慮した表現が得られる BERT (Bidirectional Encoder Representations from Transformers) [1] を用いた手法は，少ないデータで fine-tuning をすることでタスクに特化したモデルを作成でき，文書分類においても高い性能を示している．

さらに文書分類においては，著者や出版ジャーナルなどの書誌情報や引用関係にある論文，テキスト中の用語についての外部情報などの対象テキスト情報以外の情報が活用できる．Yao ら [2] は，文書と単語を節点とし，文書節点とその文書に出現する単語節点の間・共起頻度が高い単語節点間それぞれに辺を張った文書グラフの表現を用いて文書分類を行っている．この文書グラフにより，共通した単語を持つ文書節点がその単語節点を介してつながり，文書間の関係を考慮した文書分類を実現している．さらに BertGCN [3] では，文書グラフの節点表現に BERT [1] を用いてテキスト情報を追加している．また，Yasunaga ら [4] は引用関係にある文書のテキスト情報を同時に入力する事前学習手法を提案し，文書間の関係を考慮した言語モデル LinkBERT を作成し，従来の BERT より高い性能を達成している．このようにテキスト情報以外の情報の有効性は示されているものの，これらの手法では単語の共起や引用関係という限られた情報のみしか考慮できておらず，著者や出版ジャーナルなどの**書誌情報**，用語に関する説明文や上位下位関係などの外部情報（以降，**用語情報**と呼ぶ）などの複数の情報を同時に考慮できていない．

そこで，本研究では複数の情報を同時に利用できる文書分類の実現を目的に，書誌情報・用語情報など多くの対象テキスト以外の情報を含む文献グラフから文書間の様々な関係を考慮した表現ベクトルを作成し，その表現ベクトルと文書のテキスト情報を組み込んだ文書分類モデルを提案する．本研究の貢献は以下の通りである．

- 書誌情報・用語情報を含む文献グラフからの表現ベクトルを利用して，より複雑な文書間の関係の情報とテキスト情報を組み込んだ文書分類モデルを提案した．
- 生化学分野の文書分類データセットである Ohsumed [5], Hallmarks of Cancer (HoC) [6] において文献グラフの情報を用いることで既存手法からの性能向上を確認した．

2　関連研究

2.1　文書分類

文書分類には対象文書のテキスト情報のみを使用する手法 [1] と対象テキスト情報以外の情報を考慮する手法 [2, 3, 4] がある．

対象テキスト情報以外の情報を利用した手法では，文書グラフを用いた手法があり，Yao らは文書と単語を節点として，文書節点とその文書に出現する単語節点の間に TF-IDF 値で重みづけした辺を張った文書グラフを用いた文書分類を提案した [2].

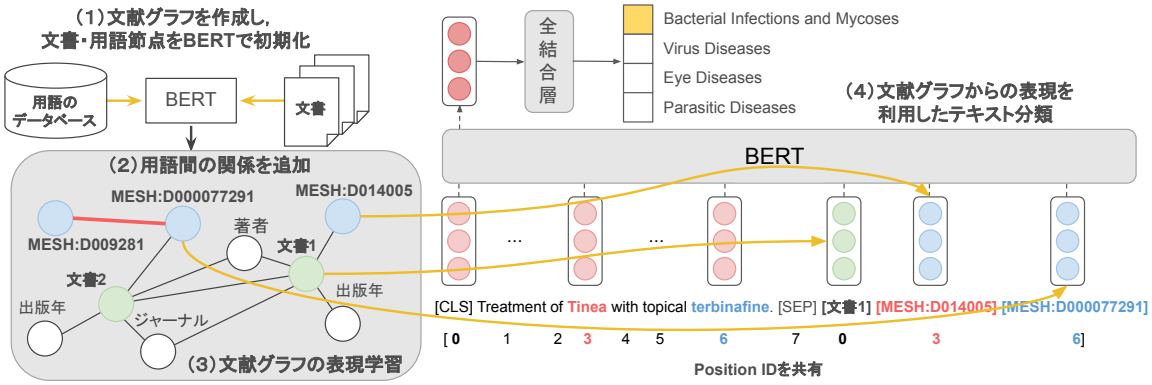

図 1　提案手法の流れ

この文書グラフでは PMI の値によって関連度が高いとみなした単語節点間にも辺を張っている．節点に接続する辺を通して周りの節点からの情報をその節点に集約するグラフ畳み込みネットワーク (Graph Convolutional Network; GCN) [7] を用いて，文書グラフの節点表現をグラフ構造を考慮しながら更新し，その表現を文書分類に利用している．さらに，BertGCN [3] では，文書節点の表現に BERT を利用し，文書グラフに文書のテキスト情報を追加する手法を提案し，高い性能を達成している．

また，Yasunaga ら [4] は引用関係などでリンクされた 2 つの文書を同時に BERT の入力とする事前学習手法により文書間の関係を考慮した事前学習モデル LinkBERT を提案した．従来の BERT で用いられる Masked Language Modeling (MLM) と，2 つの文書を入力しそれぞれが引用関係にあるかどうかを分類する Document Relation Prediction (DRP) という 2 つのタスクで BERT の事前学習をしており，文書間の依存関係や文書間にまたがる情報を利用している．その結果，LinkBERT は GLUE [8] タスクや生化学分野のベンチマークである BLURB [9] タスクにおいて既存手法の性能を上回り，文書分類においても従来の BERT より高い性能を達成している．

2.2　グラフの表現学習

文献グラフなどのグラフからグラフ構造を考慮した節点の表現ベクトルを獲得するグラフ表現学習の研究が盛んに行われている [10]．グラフ表現学習では，グラフ構造を有向辺の始点節点 h，終点節点 t，関係 r を用いて (h, r, t) と表現したトリプルの集合として表現するのが一般的であり，このトリプルの節点や関係を表現するように学習を行う TransE [11]，DistMult [12] などの手法が提案されている．TransE のスコア関数は始点節点のベクトル表現を関係のベクトル表現だけ平行移動したものと終点節点のベクトル表現との距離の大きさを損失とする．

3　提案手法

提案手法である書誌情報・用語情報を含む文献グラフの表現を組み込んだ文書分類モデルについて説明する．提案手法の概要を図 1 に示す．3.1 節で文献グラフの作成と表現学習について説明した後に 3.2 節で書誌情報・用語情報を含む文献グラフのベクトル表現と文書のテキスト情報のベクトル表現を組み込んだ文書分類モデルについて説明する．

3.1　文献グラフの作成・表現学習

まず，文献グラフを作成し，文書・用語節点を BERT で初期化する（図 1(1)）．文献グラフは文書と書誌情報を節点とし，文書節点とその書誌情報の節点間，引用関係にある文書節点間に辺を張る．文献グラフは論文・著者・出版年・出版ジャーナル・用語を節点とし，共有する節点を介して文書節点を繋げる．題目，要旨，用語の説明文がデータベースから取得できる文書・用語節点にはテキスト情報を追加する．BERT の出力の [CLS] トークンは文全体を表す表現と考えられるので，[CLS] トークンの表現を初期表現とする．データベースから取得できない場合は BERT で初期化した節点表現の平均，標準偏差の正規分布に従う乱数でランダム初期化する．書誌情報の節点についてもランダム初期化する．

次に，データベースから取得した上位下位関係，補足概念の関係にある用語節点間に辺を張り，用語間の関係も追加する（図 1(2)）．そして，作成した文献グラフについて，TransE [11][1] を用いて表現学

1）　本研究で扱う文献グラフは大規模であり，GCN [7] などの計算コストの高い手法は単純には適用できないため，簡単の

習を行い，ベクトル表現を獲得する（図 1(3)）．このように作成した文献グラフから得られたベクトル表現は対象テキスト情報以外の書誌情報・用語情報からの文書間の様々な関係を考慮したものとなっていると期待できる．

3.2 文書分類モデル

3.1 節で獲得した書誌情報・用語情報を含む文献グラフのベクトル表現と分類する文書のテキスト情報のベクトル表現を組み込んだ文書分類を行う（図 1(4)）．このために，BERT の入力として，対象文書とそれに現れる用語に対応する文献グラフの文書・用語節点のベクトル表現を文書のテキスト情報と同時に入力し，BERT で統合しながらエンコードして得られた出力の [CLS] トークンの表現を全結合層により文書のカテゴリに分類する．より具体的には，まず，文書内の用語をデータベースに登録された用語との文字列一致で抽出する．次に文献グラフの文書節点の表現と文書から抽出した用語の節点表現を文書のテキスト情報と結合する．結合の際には文書のテキスト情報の [SEP] トークンの後ろに文献グラフからの表現を追加する．この際に，テキスト内のサブワードと文献グラフの節点表現を対応付けるために，[CLS] トークンと文書節点の表現，テキスト内の用語のサブワードの先頭と用語節点の表現に同じ Position ID を割り当てる [13]．テキスト情報と対応する文献グラフ内の節点のベクトル表現を同時に BERT に入力することで，文献グラフ内の書誌情報や用語情報を活かした文書分類を目指す．

4 実験設定

文献グラフの表現学習で獲得した表現ベクトルの質を始点節点との関係ペアになる終点節点を予測するリンク予測で評価する．また，文書分類における文献グラフの表現の有効性を確かめるために文献グラフの表現を利用した文書分類の評価する．どちらの実験も BERT には BioLinkBERT-base [4] を使用する．

4.1 文献グラフの表現学習

医療文献データベース Medline [14] の 2022 年版を利用して 3.1 節で説明した文献グラフを作成する．Medline には 3,000 万件以上の論文が登録されており，文献グラフが巨大になるので他の節点とのつながりが少ない次数が 5 未満の著者節点は削除した．用語のデータベースには MeSH [15] の 2021 年版を利用した．文献グラフは引用関係を表す辺（cites），文書とその著者 (author)・出版年 (year)・出版ジャーナル (journal)・用語 (MeSH) を表す辺，用語間の上位下位関係 (hypernym)，補足概念の関係 (supp) を表す辺を持つ．付録 A の表 4 と 5 に統計を示す．

文献グラフの表現学習の評価指標には MAP@30, Hit@n を用いた．訓練・開発・評価用データはトリプルの関係タイプの比率が同じになるように 98:1:1 の割合で分割した．また，訓練用データに出現する節点間のトリプルを開発・評価用データに選んだ．リンク予測の評価の際には，訓練用データに含まれる終点節点は予測結果から削除した．また，始点節点のタイプと関係タイプから決まる終点節点のタイプの節点のみを予測対象とした．付録 B に実装に使用したライブラリと学習の設定を示した．

表 1 TransE での表現学習の結果

関係タイプ	MAP@30	Hit@1	Hit@3	Hit@10
cites	0.0046	0.0005	0.0079	0.0397
author	0.0283	0.0156	0.0335	0.1219
year	0.3261	0.1789	0.3807	0.9712
journal	0.1658	0.0973	0.1950	0.5237
MeSH	0.0870	0.0483	0.1017	0.3304
hypernym	0.0851	0.0	0.1358	0.4444
supp	0.0	0.0	0.0	0.0
マクロ平均	0.0996	0.0487	0.1221	0.3473

4.2 文献グラフの表現を利用した文書分類

評価には，医学文献の要旨の文書分類データセットである Ohsumed [5] と Hallmarks of Cancer (HoC) [6] を使用した．Ohsumed は文書に 23 種類の心血管系疾患のカテゴリのうち 1 つ以上のカテゴリが付与されている．Ohsumed の分類ラベルは MeSH から作られているので，Ohsumed に含まれる文書とその文書の書誌情報の一つである MeSH との関係が文献グラフに含まれないようにして評価した．既存研究 [2] と同様に複数ラベルを持つ文書は除外した．この結果，訓練，評価用データ内の文書はそれぞれ 3,357 件，4,043 件となった．訓練用データを 7:3 に分割して開発データを作成した．HoC ではそれぞれの文書に 10 種類の癌の特徴のカテゴリが複数付与されている．既存研究 [9] に従って訓練，開発，評価用データを分割し，それぞれ 1,295 件，186 件，371 件

ため，TransE を採用する．

表 2 文献グラフの表現を利用した文書分類の結果 [%]. 太字は最高のスコアを表す.

Method	Ohsumed	HoC
BertGCN [3]	72.8	–
PubMedBERT [9]	–	82.32
BioLinkBERT [4]	77.22	84.35
BioLinkBERT + Paper	**77.46**	**84.72**
BioLinkBERT + Entity	77.20	84.34
BioLinkBERT + Paper + Entity	76.69	84.65

を使用した．文書内の用語をデータベースに登録された用語との文字列一致で抽出した結果，Ohsumed では文書あたり平均 20.41 語，HoC では文書あたり平均 26.56 語の用語を抽出できた．

評価においては，乱数シードを変更した 5 つのモデルを作成し，その評価の平均を最終的な予測結果として報告する．Ohsumed の評価には正解率，HoC の評価には F 値を用いた．ベースラインモデルとしては，3.2 節の文書分類モデルの入力にテキスト情報のみを利用するモデル (BioLinkBERT) を用意した．また，文書分類モデルに文献グラフから追加する節点情報としては，文書の表現 (+ Paper)，用語の表現 (+ Entity)，文書・用語の両方の表現 (+ Paper + Entity) の 3 つの組み合わせを比較をした．付録 B に実装に使用したライブラリと学習の設定を示した．

5 結果と考察

5.1 結果

TransE での表現学習の結果を表 1 に示す．今回使用した文献グラフは節点数が多かったため，MAP@30, Hit@n ともに全体的に低い値となった．特に引用関係，MeSH 間の関係の性能が低いのは，一つの始点節点との関係ペアになる終点節点が複数あるため，TransE では表現できなかったためだと考えられる．

文献グラフの表現を利用した文書分類の結果を表 2 に示す．Ohsumed における BioLinkBERT の結果と提案した情報を追加したモデル (BioLinkBERT + Paper，BioLinkBERT + Entity，BioLinkBERT + Paper + Entity) 以外の値は元論文から引用したものである．Ohsumed については，これまでに報告されている分類スコアを大きく上回ることができた．また，論文の表現のみを使用したときに Ohsumed, HoC ともにベースライン (BioLinkBERT) からの性能向上が見られた．一方で，用語の表現のみを使用したときにはほとんど性能の違いが見られなかった．論文・用語の表現の両方を使用したときには Ohsumed では性能低下，HoC では僅かな性能向上が見られ，一貫性のない結果となった．いずれの表現も文献グラフ全体を考慮して学習された表現ではあるものの，使う表現によって結果に違いがあることがわかった．

表 3 用語をタグで置き換えた文書分類の結果 [%]

Method	HoC
BioLinkBERT [4]	79.85
BioLinkBERT + Paper	**81.26**
BioLinkBERT + Entity	80.94
BioLinkBERT + Paper + Entity	80.99

5.2 考察

文献グラフの表現ベクトルの質を確かめるために，対象テキスト情報から抽出した用語をマスクして実験をした．その結果を表 3 に示す．対象テキストの用語をマスクしたときベースラインモデルでは，4.5 ポイント性能が低下した．一方で，文書の表現のみ，用語の表現のみ，文書・用語の両方の表現を使用したときはそれぞれ 3.46 ポイント，3.40 ポイント，3.66 ポイントの性能低下となり，ベースラインよりも性能の低下が小さかった．この結果から，提案手法では対象テキスト情報以外の情報を分類に利用できていることが分かった．付録 C にテキスト情報を利用しない文書分類の結果と考察を示した．

6 おわりに

本研究では複数の情報を同時に利用できる文書分類を目的として，書誌情報・用語情報など多くの情報を含む文献グラフから文書間の様々な関係を考慮した表現ベクトルを作成し，その表現ベクトルと文書のテキスト情報を組み込んだ文書分類モデルを提案した．Ohsumed, HoC の 2 つのデータセットで実験を行った結果，文書の表現のみを利用したときにはどちらも性能向上が見られた．特に，Ohsumed では従来モデルを上回る性能を達成した．論文・用語の表現の双方を利用すると性能が低下することがあるという現象がみられ，異種の情報の組み込みが単純ではないことを示すことができた．

今後は，このような問題を解決するために GCN などの手法での大規模文献グラフの表現学習を検討する．また文献グラフの表現学習手法と文書分類モデルを同時に学習する方法についても模索する．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです.

参考文献

[1] Jacob Devlin, et al. 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.

[2] Liang Yao, et al. Graph convolutional networks for text classification. In **Proceedings of the AAAI conference on artificial intelligence**, Vol. 33, pp. 7370–7377, 2019.

[3] Yuxiao Lin, et al. BertGCN: Transductive text classification by combining GNN and BERT. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 1456–1462, Online, August 2021. Association for Computational Linguistics.

[4] Michihiro Yasunaga, et al. Linkbert: Pretraining language models with document links. In **Association for Computational Linguistics (ACL)**, 2022.

[5] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In **Proceedings of the 10th European Conference on Machine Learning**, ECML'98, p. 137–142, Berlin, Heidelberg, 1998. Springer-Verlag.

[6] Simon Baker, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Vol. 32 3, pp. 432–40, 2016.

[7] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In **International Conference on Learning Representations**, 2017.

[8] Alex Wang, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[9] Yu Gu, et al. Domain-specific language model pretraining for biomedical natural language processing. **ACM Transactions on Computing for Healthcare**, Vol. 3, No. 1, oct 2021.

[10] William L Hamilton. Graph representation learning. **Synthesis Lectures on Artifical Intelligence and Machine Learning**, Vol. 14, No. 3, pp. 1–159, 2020.

[11] Antoine Bordes, et al. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, **Advances in Neural Information Processing Systems**, Vol. 26. Curran Associates, Inc., 2013.

[12] Bishan Yang, et al. Embedding entities and relations for learning and inference in knowledge bases. In **International Conference on Learning Representations**, 2014.

[13] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 50–61, Online, June 2021. Association for Computational Linguistics.

[14] Pubmed. https://pubmed.ncbi.nlm.nih.gov/（1 月 2 日アクセス）.

[15] Medical subject headings - home page. https://www.nlm.nih.gov/mesh/meshhome.html（1 月 2 日アクセス）.

[16] Da Zheng, et al. Dgl-ke: Training knowledge graph embeddings at scale. In **Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '20, p. 739–748, New York, NY, USA, 2020. Association for Computing Machinery.

[17] Thomas Wolf, et al. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

[18] Adam Paszke, et al. Pytorch: An imperative style, high-performance deep learning library. **Advances in neural information processing systems**, Vol. 32, pp. 8026–8037, 2019.

[19] Masajiro Iwasaki and Daisuke Miyazaki. Optimization of indexing based on k-nearest neighbor graph for proximity search in high-dimensional data. **CoRR**, Vol. abs/1810.07355, , 2018.

[20] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, Vol. 15, No. 1, pp. 1929–1958, 2014.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **International Conference on Learning Representations**, 2019.

A　グラフの統計

文献グラフの統計を表 4 と 5 に示す．テキスト情報を追加した節点数は，文書節点は 33,404,632 個，用語節点は 244,373 個である．

表 4　文献グラフの節点の統計

節点タイプ	
文書	33,406,096
著者	4,932,150
年	57
出版ジャーナル	34,564
MeSH	348,081
Total	38,720,948

表 5　文献グラフのトリプル統計

関係タイプ	全体	Train	Valid	Test
cites	246,136,539	241,213,809	2,461,365	2,461,365
author	118,193,406	115,829,538	1,181,934	1,181,934
year	33,405,863	32,737,747	334,058	334,058
journal	33,405,863	32,737,747	334,058	334,058
MeSH	31,917,346	31,279,000	319,173	319,173
hypernym	40,659	39,847	406	406
supp	427,758	419,204	4,277	4,277
Total	463,527,434	454,256,892	4,635,271	4,635,271

B　学習の設定

実装には，Python 3.7.11 を使い，TransE には DGL-KE 0.1.2 [16]，事前学習モデルを使うために Transformers 4.19.4 [17]，モデルの作成のために Pytorch 1.10.0 [18] を用いた．また，リンク予測の評価は，近傍探索ライブラリ NGT [19] を用いて近似的に求めた．TransE は 50 エポック学習した．文献グラフの文書・用語節点の表現は BERT で初期化するため表現は 768 次元となるので，ランダム初期化する節点の表現についても 768 次元で初期化した．TransE の学習ではグラフに含まれるトリプル (h, r, t) の h か t のどちらかをランダムに置き換えるネガティブサンプリングが使用される．始点節点のタイプと関係タイプから決まる終点節点のタイプのみからネガティブサンプリングをした．TransE 学習のための計算機には CPU に AMD Ryzen Threadripper 3990X 64-Core Processor，GPU に GeForce RTX 3090 を使用した．

文書分類モデルでは，BERT の CLS トークンの表現を一層の線形層で分類している．また，訓練用データでの過学習を防ぐために線形層の前にドロップアウト [20] を加えた．文書分類学習のための計算機には CPU に Intel(R) Xeon(R) CPU E5-2698 v4 及び Intel(R) Core(TM) i9-10900K CPU，GPU には Tesla V100-DGXS-32GB 及び GeForce RTX 3090 を使用した．文書分類モデルの最適化手法には AdamW [21] を使用した．学習率は {1e-5, 2e-5, 3e-5, 4e-5, 5e-5, 6e-5} の範囲でハイパーパラメータチューニングを行い，表 6 のように設定した．

表 6　ハイパーパラメータチューニングの結果

Method	Ohsumed	HoC
BioLinkBERT	4e-5	4e-5
BioLinkBERT + Paper	4e-5	6e-5
BioLinkBERT + Entity	4e-5	5e-5
BioLinkBERT + Paper + Entity	4e-5	5e-5

C　テキスト情報を除いた文書分類

文献グラフから得られた表現単体での性能を評価するため，対象テキスト情報を使用しない実験を行った．BERT への入力を文献グラフからの表現のみにして出力の [CLS] トークンで分類をした結果を表 7 に示す．文書・用語の表現の両方を使用したときに最も高い性能となった．様々な情報を含む文献グラフからの表現を利用することで対象テキスト情報がなくても分類できたと考えられる．

表 7　テキスト情報を除いた文書分類の結果 [%]

Method	HoC
BioLinkBERT + Paper	39.64
BioLinkBERT + Entity	64.32
BioLinkBERT + Paper + Entity	**66.23**