説明可能な検索ベースの文書分類手法の提案

中井優 [1]　中野雄介 [1]　徳永優也 [1]　上田亮 [1]　谷中瞳 [1]
[1] 東京大学
{nakai-yu623,nakano-yusuke,yn-noob}@g.ecc.u-tokyo.ac.jp
{ryoryoueda,hyanaka}@is.s.u-tokyo.ac.jp

概要

現在，文書分類などの様々な自然言語処理のタスクにおいて，判断根拠が説明可能な手法の重要性が高まっている．本研究では，分類に有効な文書を訓練事例や外部の知識コーパスから検索して利用する，検索ベースの文書分類手法を提案する．提案手法は，分類に利用した文書を出力することができ，文書分類に説明可能性を与える．本論文では，AG News データセットを用いて学習・推論を行い，分類精度・説明可能性の 2 つの側面から提案手法の有効性を検証した．実験の結果，知識コーパスを利用しない既存の大規模言語モデルと比べて，知識コーパスの検索結果を利用することによって精度が向上することを確認した．また，モデルが分類根拠として示す文書と分類対象の文書との文間類似度と，人手評価との間に有意な相関を確認することができた．

1　はじめに

自然言語処理のタスクに文書分類がある．文書分類とは，与えられた文書に対して事前に定義されたラベル群から適切なラベルを推定するタスクのことである．文書分類を自動化するための技術は実社会でも広く活用されており，その社会実装に対するニーズも相まり，近年文書分類のためのニューラルネットワーク (NN) モデルが盛んに研究されている．

他方で，NN モデルの推論機構はブラックボックス化されており，結果に対する判断根拠の不透明さが社会実装を妨げる課題となっている．このような不透明性を軽減するための一連の技術が，**説明可能 AI (XAI)** という研究トピックを形成している．

説明可能 AI の一分野として，**事例ベース XAI** と呼ばれるものがある [1, 2, 3]．事例ベース XAI とは，評価データに対する NN モデルの予測において，その予測に最も影響を与えたであろう訓練事例を探ることで，予測における判断根拠を説明しようとするアプローチである．

本研究では，分類に有効な文書を訓練事例や外部の知識コーパスから検索して利用する，**検索ベースの文書分類手法**を新たに提案する．提案手法は，分類に利用した文書を出力することにより分類の根拠に説明可能性を与える．この検索ベースの手法は，訓練事例や外部の知識コーパスを直接参照することで文書分類の性能向上を図る手法であり，事例ベース XAI とは訓練事例の扱い方が異なるものの，文書分類の根拠が説明可能であるという点で類似するアプローチであるといえる．

提案する NN モデルは Retriever と Classifier の 2 つのネットワークより構成される．Retriever は分類対象の文書に対して知識コーパス（以降では，検索対象の文書を総称して知識コーパスと呼ぶ）からその分類に有意な文書を検索・抽出し，Classifier は分類対象の文書と抽出した文書を用いて分類を行う．この新たに提案する検索ベースの文書分類モデルを**検索ベース分類器 (Retrieval-based Classifier)** と呼ぶ．

本論文では，ニュース記事の分類タスクのデータセットである AG News データセットを用いて学習・推論を行い，(i) 分類精度と (ii) 説明可能性という 2 つの側面から，提案手法の有効性を検証する．実験の結果，知識コーパスを利用しない既存の大規模言語モデルと比べて，知識コーパスの検索結果を利用することによって精度が向上することを確認した．また，モデルが分類根拠として示す文書と分類対象の文書の文間類似度と，人手評価との間にも有意な相関を確認することができた．

2　関連研究

自然言語処理の分野において，これまで説明可能 AI に関する様々な研究が行われてきた．以下ではその中での本研究の位置付けを明らかにするとともに，提案手法のベースになるモデルである検索拡張言語モデルについて説明する．

2.1 自然言語処理における説明可能 AI

[4] によれば，説明可能 AI における説明の種類は，大きく分けて次の 2 つに分類される：

- Post-Hoc：モデルの予測後に追加の操作を行って予測を説明する
- Self-Explaining：モデルの予測過程から得られる情報によって予測を説明する

提案手法はモデルの予測過程で利用した文書を出力することによって文書分類の判断根拠が説明可能であることから，Self-Explaining な説明可能 AI として位置付けられる．

自然言語処理における Self-Explaining な説明可能 AI としては，Attention や First-derivative saliency などの特徴量の重要度から説明を導出する方法 [5, 6] が提案されている．これに対して，提案手法は分類に利用した文書を予測過程から出力できるため，既存手法よりも直接的に文書分類の判断根拠を提示する手法である．

2.2 事例ベース XAI

XAI の一分野として，事例ベース XAI というものがある．事例ベース XAI とは，評価データに対する NN モデルの予測において，その予測に最も影響を与えたであろう訓練事例を探索することで，予測における判断根拠を説明しようとするアプローチである．

事例ベース XAI のアプローチから着想を得て，我々は，検索ベースの文書分類手法を XAI として活用できないかと考えた．すなわち，モデルが分類時に選択した知識コーパスの文書が暗に分類時の判断根拠となっているとみなせば，選択した文書を出力することでモデルの判断根拠を説明できる可能性がある．ただし，我々の提案手法は Self-Explaining なものであるのに対し，多くの事例ベース XAI [1, 2, 3] は，ブラックボックスな NN モデルの推論結果に後から説明可能性を与えるという意味で Post-Hoc [4] な手法として位置付けられる点に相違がある．本研究では，分類時に選択した知識コーパスの文書の説明可能性について，実験を通して検証を行う．

2.3 検索拡張言語モデル

ドメインが指定されていない質疑応答タスクである Open-Domain Question Answering の分野においては，検索拡張言語モデル (Retrieval-Augmented Language Models) [7, 8, 9, 10] が広く用いられている．検索拡張言語モデルは通常 Retriever と Reader の 2 つのネットワークによって構成される．Retriever は入力の質問文に対して外部の知識コーパスからその解答の作成に有意な文書を検索・抽出し，Reader は質問文と抽出した文書を用いて回答文を作成する．

提案した検索ベース分類器は，この Retriever と Reader からなる構成から着想を得た．

3 提案手法

本研究で提案するモデルである検索ベース分類器は，図 1 で示すように知識コーパスから分類対象の文書に関連する文書を抽出する Retriever と，分類対象の文書と抽出された文書から文書分類を行う Classifier の 2 つのネットワークによって構成される．

分類対象の文書 x，外部の知識コーパス z，分類クラスのラベル y として，分類時の計算手順は次の数式で表される．

$$p(y|x) \propto \sum_{z \in \mathrm{Top_k}(p(\cdot|x))} p_\mu(z|x) p_\theta(y|x,z) \qquad (1)$$

ただし p_μ と p_θ はそれぞれ，次に説明する Retriever と Classifier を指す．

3.1 Retriever

Retriever は，知識コーパスの埋め込みを行う Document Encoder，分類対象の文書の埋め込みを行う Query Encoder，及び分類対象の文書と知識コーパス中の文書との最大内積探索 (MIPS) から構成される．ここで最大内積探索では，知識コーパス中の文書の埋め込み表現のうち，分類対象の文書の埋め込み表現とのコサイン類似度が高いものを探索する．

Document Encoder d と Query Encoder q に対して，知識コーパス中の文書 z の尤度を次式によって算出する．

$$p_\mu(z|x) = \frac{d(z)^T q(x)}{\|d(z)\|^2 \|q(x)\|^2} \qquad (2)$$

この式の $p_\mu(z|x)$ が，分類対象の文書 x に対する，知識コーパス中の文書 z の文間類似度を示す．

3.2 Classifier

Classifier の計算手順は，文書分類器 p を用いて次式によって表される．ここで，|| は系列の結合を行

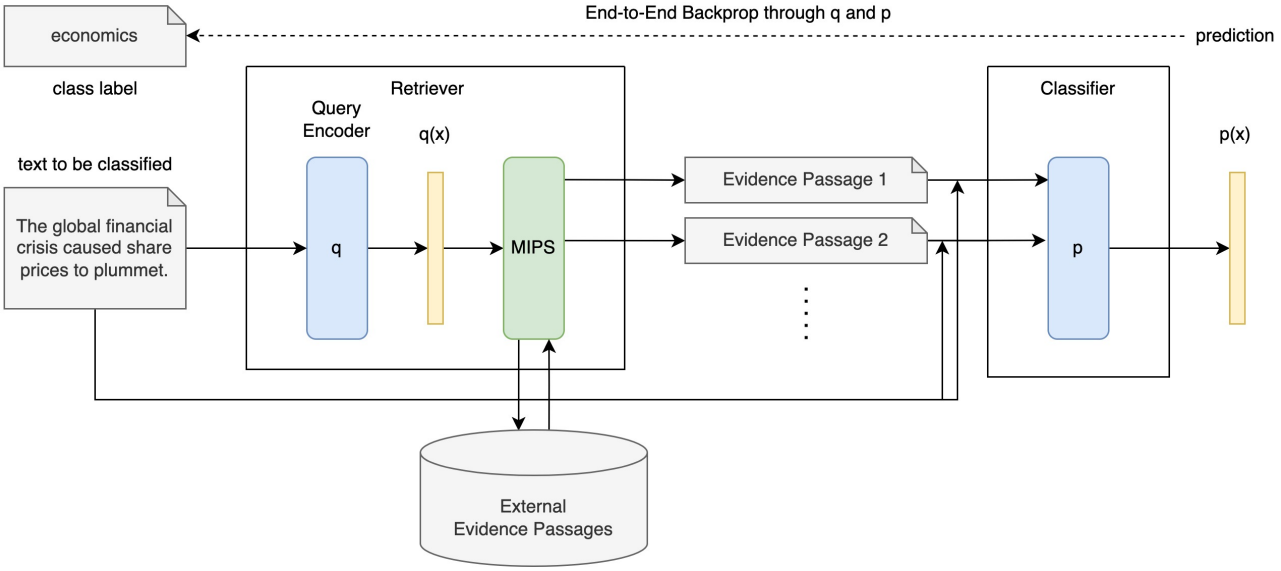

図 1　検索ベース分類器の構造. Retriever は Query Encoder と最大内積探索 (MIPS) を用いて知識コーパスから文書の抽出 (図中の external evidence passage) を行い，Classifier は分類対象の文書と文書を用いて分類を行う.

う演算子を指す.

$$p_\theta(y|x,z) = p(x||z) \tag{3}$$

本研究では文書分類器 p として，大規模言語モデル BERT [11] と 1 層の全結合層からなる構造を利用した.

4　実験

4.1　実験設定

データセット　判断根拠提示の実用上の意義の観点から，データセットとしては英文ニュース記事の分類タスクである AG News データセット [12] を利用した．AG News は，約 16 万件の英文ニュース記事とそのジャンルに関するラベルを有するデータセットである.

評価指標　分類性能の評価指標として，Accuracy を採用した．説明可能性の判断指標として，異なる 2 文の類似度を評価するための指標である意味類似度 (STS) [13] を採用した．ここでは情報系の大学院生 3 名 (いずれも日本語話者) が STS [13] の定義に基づき 5 段階で評価を行い，その平均を取った．なお本アノテーションにおいては報酬などは用意せずに実施した.

実験設定　AG News の 1,600 件のニュース記事を訓練データ，158,400 件のニュース記事のうちの一部を知識コーパス，7,600 件のニュース記事を評価用データとして利用し，次の 2 つの実験を行った.

1 つ目に，Retriever の精度に与える影響を評価することを目的として，(i) Classifier のみによる文書分類，(ii) 予測過程にて訓練事例のみを利用した文書分類，(iii) 予測過程にて訓練事例と知識コーパスを利用した文書分類の 3 通りについて，ランダムシード値を変えてそれぞれ 10 回ずつ実験を行った．2 つ目に，Retriever が出力する文書の説明可能性を評価することを目的として，評価用データからサンプリングした 40 件のニュース記事について Retriever を用いて関連文書との文間類似度を出力し，その文間類似度と人手評価による意味類似度との相関を分析する実験を行った．文間類似度について満遍なく実験を行うため，サンプリングでは Retriever の出力する文間類似度が 1,10,100,1000 番目に高い関連文書を抽出した.

学習条件　Retriever には，Query Encoder と Document Encoder のそれぞれの BERT について教師なし対照学習により事前学習を行ったモデルである，Contriever [14] の学習済み重みを初期重みとして利用した．Document Encoder により訓練データ及び知識コーパス中の文書の埋め込みを行い，最大内積探索用ライブラリである faiss [15] のインデックスとして追加した．Classifier は BERT [11] の学習済み重みを初期重みとして利用した．全結合層は Xavier の重みの初期化 [16] を行なった．訓練データを用いた Retrieval-based Classifier の学習時には，Document Encoder の重みとインデックスは固定し，Query Encoder と Classifier のファインチューニング

表 1 (i) Classifier のみによる文書分類，(ii) 予測過程にて訓練事例のみを利用した文書分類，(iii) 予測過程にて訓練事例と知識コーパスを利用した文書分類の，AG News における分類精度の平均と標準誤差．どの場合も，AG News の 1%のみを学習データとして利用した．

モデル	Accuracy
(i) Classifier のみ	87.43 ± 08
(ii) 訓練事例のみを利用	87.31 ± 25
(iii) 訓練事例と知識コーパスを利用	87.63 ± 07

を行なった．また，最適化関数には Adam [17] を利用し，損失関数としては交差エントロピー損失を利用した．

4.2 実験結果

4.2.1 分類精度の比較評価

(i) Classifier のみによる文書分類，(ii) 予測過程にて訓練事例のみを利用した文書分類，(iii) 予測過程にて訓練事例と知識コーパスを利用した文書分類のそれぞれの場合の AG News の分類精度を表 1 に示す．

この結果から，Retriever の導入により分類精度が向上することを確認できた．知識コーパスを利用せずに訓練事例のみを用いた場合には，Classifier のみの場合と比較して若干精度が落ちてしまっているものの，訓練事例に加えて知識コーパスまで含めて検索対象にすることで，Classifier のみの場合よりも高い精度を達成できた．このことから，適切な量の知識コーパスのもとで，Retriever は分類精度の向上に寄与する可能性があると推察される．

なお，(iii) 予測過程にて訓練事例と知識コーパスを利用した文書分類では事前に適切な知識コーパスの量の探索を行っており，知識コーパスが訓練事例数の 3 倍の文書数を持つときに最も高い性能を発揮した．これらの知識コーパスの量に対する分類精度の変化に関する結果は付録 A に記載した．

4.2.2 説明可能性の評価

Retriever が分類根拠として示す文書と分類対象の文書との間の，人手評価に基づく意味類似度と Retriever の示す文間類似度との関係を散布図に示したのが図 2 である．相関係数は 0.61 ($p = 3.0 \times 10^{-5}$) であり，人手による意味類似度と Retriever の示す文間類似度には有意な相関があることを示している．

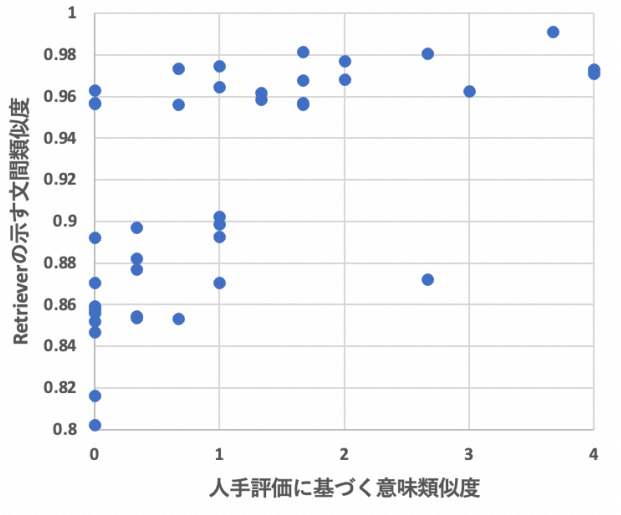

図 2 Retriever が分類根拠として示す文書と分類対象の文書との間の，人手評価に基づく意味類似度と Retriever の示す文間類似度との関係を表す散布図．

この結果は Retriever が分類時に利用した文書は人間から見ても分類対象の文書と有意に関連のある文書であることを示しており，提案手法が説明可能性を有することを裏付ける．

なお，Retriever が実際に分類根拠として示した文書の例を付録 B に記載した．

5 結論

本研究では，分類に有効な文書を訓練事例や外部の知識コーパスから検索して利用する，検索ベースの文書分類手法 (Retrieval-based Classifier) を提案した．提案手法の性能を評価するために，AG News データセットを対象とし，Retriever や知識コーパスの存在が分類精度にどのように影響するのかを評価した．また，事例ベース XAI に関する先行研究から着想を得て，提案手法における Retriever が選択した根拠文が説明可能性に寄与する可能性に着目し，それに関する定性的な評価を行なった．一連の実験の結果から，文書分類モデルに Retriever の機構を導入することで，精度を落とすことなく説明可能性を与えられることが確認できた．ただし，今回は特定のデータセットを対象として，限られた人数での定性評価をするにとどまっている．適用するドメインやデータセットの種類を増やしていくとともに，より信頼性の高い定性評価を行うことで，我々の提案手法の有用性を明瞭にしていくことが今後の課題である．

謝辞

本研究は，JSPS 科研費 JP20K19868，JST さきがけ JPMJPR21C8 の助成を受けたものです．

参考文献

[1] Chih-Kuan Yeh, Joon Sik Kim, Ian En-Hsu Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks. In **Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada**, pp. 9311–9321, 2018.

[2] Rajiv Khanna, Been Kim, Joydeep Ghosh, and Sanmi Koyejo. Interpreting black box predictions using fisher kernels. In **The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan**, Vol. 89 of **Proceedings of Machine Learning Research**, pp. 3382–3390. PMLR, 2019.

[3] Amirata Ghorbani and James Y. Zou. Data shapley: Equitable valuation of data for machine learning. In **Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA**, Vol. 97 of **Proceedings of Machine Learning Research**, pp. 2242–2251. PMLR, 2019.

[4] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In **Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing**, pp. 447–459, Suzhou, China, December 2020. Association for Computational Linguistics.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

[6] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 681–691, San Diego, California, June 2016. 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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers**, pp. 6086–6096. Association for Computational Linguistics, 2019.

[8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In **Proceedings of the 37th International Conference on Machine Learning**, ICML'20. JMLR.org, 2020.

[9] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.

[10] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022. https://arxiv.org/abs/2208.03299v3.

[11] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[12] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In **Advances in Neural Information Processing Systems**, Vol. 28. Curran Associates, Inc., 2015.

[13] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[14] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Towards unsupervised dense information retrieval with contrastive learning. **CoRR**, Vol. abs/2112.09118, , 2021.

[15] J. Johnson, M. Douze, and H. Jegou. Billion-scale similarity search with gpus. **IEEE Transactions on Big Data**, Vol. 7, No. 03, pp. 535–547, jul 2021.

[16] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In **Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010**, Vol. 9 of **JMLR Proceedings**, pp. 249–256. JMLR.org, 2010.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

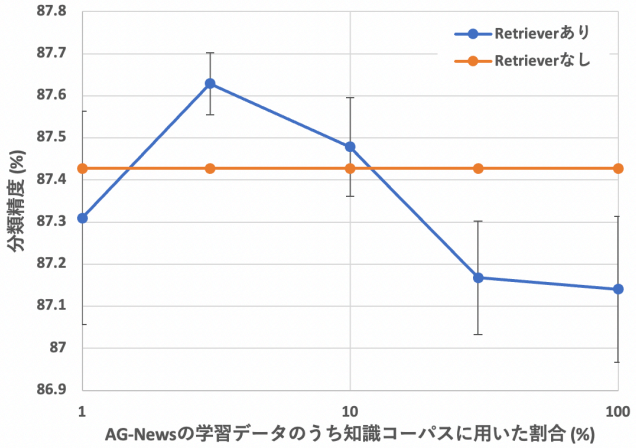

図 3　Retriever にて用いる知識コーパスの量に対する分類精度の変化.

表 2　分類対象の文書と Retriever が抽出したエビデンス文書

分類対象の文書	抽出したエビデンス文書
iran may soon resume uranium enrichment (ap) ap - iran may resume uranium enrichment any moment, the nation’ s intelligence minister said on state television monday, two days after the u. n. nuclear watchdog agency demanded that tehran halt all such activity.	iran stresses nuclear freeze is voluntary, brief tehran (reuters) - iran stressed on monday its decision to freeze sensitive nuclear work was a voluntary move to dispel concerns it was secretly building atomic arms and would last only for a short time.
japan’s smfg in $ 29b bid for ufj sumitomo mitsui financial group inc. laid out a $ 29 billion bid for ufj holdings on tuesday, challenging a rival offer by mitsubishi tokyo financial group to form the world’s biggest bank.	japan’s sumitomo tables ufj bid the battle to form the world’s biggest bank has intensified after sumitomo mitsui financial group tabled an offer to buy japanese rival ufj holdings.

A　知識コーパスの量に対する分類精度の変化

図 3 は，Retriever で用いる知識コーパスの量を AG News の 1%から 100%まで変化させたときの分類精度の変化を示す.

この結果から，検索対象の知識コーパスの量が少なすぎる場合や多すぎる場合には，むしろ Retriever の導入は精度を悪化させる可能性があり，適切な量の知識コーパス (今回であれば訓練事例数の 3〜10 倍の文書数) を利用することで精度向上に寄与する可能性があるということが推察される.

B　Retriever が抽出した文書の例

表 2 には，分類対象の文書と知識コーパスより Retriever が抽出した文書のペアの一例を示す．この例から分かるように，ある程度関連性のある文書を分類の根拠として出力することができているといえる．例の 1 つ目はイランの核開発問題，2 つ目は日本のフィナンシャルグループの買収に関する記事である.