金融文書の抽象型要約による投資家向け支援システムの提案

中野凌[1] 蓮池隆[2]
[1]早稲田大学大学院 [2]早稲田大学 創造理工学研究科
ryoku.nkn.18@toki.waseda.jp thasuike@waseda.jp

概要

投資家は，決算短信，有価証券報告書，株主招集通知，アナリストレポートといった金融文書を元に投資先を決定するが，金融文書の量は膨大であり，それらの分析には多大は労力と専門的な知識を要する．そこで本研究では，金融文書に対する自動要約技術の適用に着目し，様々な要約モデルを比較することで，実用性の有無を検証した．

本研究の貢献は以下の 3 つである．

1. Web スクレイピングによる要約タスクのファインチューニング用データセットの収集を行った．
2. 収集したデータセットを様々な条件で抽出，ファインチューニングすることで日本語版 BART モデルの要約タスクにおける精度を定量的に評価した．
3. BART モデルと他の自動要約手法の定量的・定性的な評価によって比較し，金融文書の抽象型要約による投資家向け支援システムを提案した．

1 研究背景・目的

1.1 自動要約技術について

自動要約は，入力された文章を短くまとめて自動的に出力する技術である．近年，インターネットの普及や文書の電子化などにより，テキスト媒体におけるデータが大量に蓄積されているが，その中には重要性の低い文章も含まれている．そこで自動要約により重要な箇所のみを抽出することで，人間は文章を読む時間を削減でき，より本質的な内容を吟味することができる．社会における自動要約の活用場所として，広告や雑誌，求人などの文字数制限のある媒体に対する見出し作成や金融文書や法律文書などからの重要箇所の取得などがあげられる．これらの背景から，文章要約は自然言語処理のタスクの中でも重要度が増している．本研究では，金融文書に対する自動要約に着目し，様々な文章要約モデルを比較することで，実用性の有無を検証することを目的とする．

1.2 一億総株主の課題

日本の個人金融資産構成は長年にわたって預貯金に偏っており，欧米諸国に比べて非効率な運用が常態化している．そこで 2022 年，政府は資金シフトを促すため，「一億総株主」といわれるすべての国民が日本企業への投資家となることを目指す施策を講じた．これにより，今後，投資を始める人が増加することが予想される．投資家は，決算短信，有価証券報告書，株主招集通知，アナリストレポートといった金融文書を元に投資先を決定するが，金融文書の量は膨大であり，それらの分析には多大は労力と専門的な知識を要する．そのため，投資初心者が大量の金融情報から投資先を選択することはハードルが高く，「一億総株主」に向けての課題点となっている．そこで，本研究では，企業の金融文書を収集し，自動要約することで，投資初心者にも分かりやすい金融情報を提供する手法を提案することを目的とする．

2 関連研究と本研究のアプローチ

2.1 金融文書の要約

文章要約には，抽出型要約と抽象型要約の 2 種類がある．抽出型要約は，入力した文章から重要な文章を抜き出して要約文を生成する手法である．抽象型要約は，入力した文章を元に一から文章を生成する手法である．金融文書に対する抽出型要約の活用として，平野ら[1]は，アノテーションの必要のない別のタスク(銘柄判定，価格変動)を学習させた学習モデルの Attention を用いることで，教師なしで金融文書の重要文判定を行う手法を提案している．抽出型要約は，確実に元の文章内の正しい内容が出力されるというメリットがあるが，デメリットとして，文と文の接続に違和感がある，要約文の長さを制御しにくいなどの点がある．そこで，本研究では抽象型要約を用いて金融文書を要約することで違和感がなく読みやすい文章を生成する手法を提案する．

2.2 BART

Bidirectional and AutoRegressive Transformers[2](以下，BART)は Bidirectional Encoder Representations from Transformers[3](以下，BERT)と Generative Pretrained Transformer[4](以下，GPT)の両方のアーキテクチャを持ち，BERT を Encoder として入力処理し，エンコーディングされた情報を使用し GPT の Auto-regressive Decoder により文書を生成するモデルである．Katsumata ら[5]は BART を用いて英語の文法誤り訂正タスクを行っており，既存の手法で最もよい結果が得られたと報告している．また，田中ら[6]は，約 1800 万文の日本語 Wikipedia を用いて学習を行った日本語版 BART の事前学習モデルを構築し，日本語の入力誤り認識タスクにおいて，他の校正支援 API と比較して，精度が高いことを確認して

いる．しかしながら，日本語版 BART による要約タスクの精度を比較した論文はみられない．そこで，本研究では，田中らの作成した日本語版 BART の事前学習モデルに対して，文章要約タスクのファインチューニングを行い，要約精度の比較を行う．

3 提案手法

本研究の提案手法の全体像を図 1 に示す．提案手法は 4 つのステップから構成される．

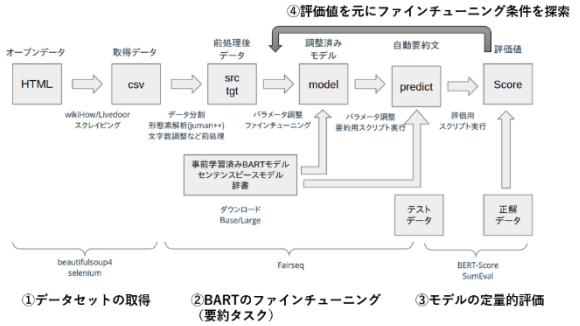

図 1. 提案手法の概要図

3.1 要約用データセットの収集

BART の文章要約タスクのファインチューニングのため，要約前の文章と要約後の文章がペアになったデータセットを構築する必要がある．本研究では，wikiHow[7]と livedoor[8]の 2 つのサイトから Web スクレイピングによってデータセットを収集した．

wikiHow はアート・健康・趣味など様々な分野に関するハウツーが記載されたサイトである．ハウツーは「タイトル」「方法」「ステップ」から構成され，各見出し文を合わせたものを要約後の文章，その他の文章を要約前の文章として Web スクレイピングを行い，データセットを構築した．

livedoor は政治，経済，IT，スポーツなど様々なジャンルのニュース記事を掲載するサイトである．記事ごとに「3 行要約ページ」と「記事詳細ページ」が存在するため，それらを用いてデータセットを構築した．ThreeLineSummaryDataset[9]を利用した文章要約データセットの取得により，2013 年 1 月～2016 年 12 月までに公開された記事を取得した．また，本研究ではより多くの学習データを収集するため，2017 年 1 月～022 年 12 月までのデータセットについてもニュース ID を探索することで，新たなデータセットの構築を行った．結果として，以下の件数の要約データセットを収集した．

- wikihow：35,374 件
- livedoor2013 年～2016 年：101,559 件
- livedoor2017 年～2022 年：119,741 件
- 上記の全データ合計：256,674 件

3.2 BART のファインチューニング

まず，3.1 節で収集したデータセットの前処理として，要約前の文字数が 1500 文字以内の条件でデータを抽出した．これは，日本語版 BART の事前学習モデルにおいて，token の長さが 1024 を超えるデータは処理できないためである．次に，抽出後のデータに対して，Juman++[10]による形態素解析とセンテンスピースモデル[11]を適用し，前処理済みデータを作成した．最後に，前処理済みデータを train，test，validataion データにおよそ 8:1:1 の割合で分割し，ファインチューニングを行った．

3.3 自動要約結果とモデルの定量的評価

3.2 節のファインチューニング済みモデルを用いて，自動要約を行った．表 1 に例として， livedoor のニュース記事の文章を日本語 BART による自動要約した結果と livedoor の 3 行要約ページの正解の要約文を示す．要約前の文章は，約 1500 文字になるため，文字数の都合上省略する．

表 1. 日本語版 BART 抽象型要約の要約例

日本語BARTの抽象型要約文	正解要約文
市の情報通信会社が６月から、仮想現実（ＶＲ）で挙式やフラワーシャワーを体験できる「すぐ婚ＶＲ」を発売した。当事者の目線で式を体験するシステムとしては全国初という。ただ映像を見ているだけのような形ではなく、没入感、臨場感を出せるようこだわった。	名古屋市の会社が6月から、「すぐ婚VR」を開発し提供している仮想現実（VR）で結婚式を疑似体験できるサービスとなっている20代男性エンジニアの提案から始まり、没入感などを出せるようしたという

表 1 より，抽象型要約で違和感のない文章が生成されていることが分かる．

次に，ファインチューニング済みモデルの評価のため，ファインチューニングに用いなかったテスト用の 108 のデータセットを用いて定量的評価を行った．評価では，自動要約によって生成した文章と正解の要約文を比較することで定量的なスコアを算出する．本研究では，評価手法として，以下の 4 種類を用いた．

- BERTScore：事前学習された BERT から得られるベクトル表現を利用して，文書間の類似度を計算する．
- BLUE : 生成した要約の N-gram 中のどれだけが正解とする要約に登場するかを計算する．
- ROUGE-N：2 つの文書館の N-gram 単位での一致度を評価する．
- ROUGE-BE：文法的な要素間の係り受けを考慮し評価する．

3.4 ファインチューニング条件の探索

3.2 節で行ったファインチューニングにおいて，3.3 節の定量的評価が最もよくなるような条件を探索した．条件は以下の 5 つである．

- 事前学習 BART モデル：base，large
- ファインチューニング学習回数 : 3，5，10epoch
- 使用データセット：wikiHow，Libedoor，両方
- 文字数：300~1000，0~1500，300~1500 文字
- 文字数圧縮率：10~30，5~50%

表 2. 日本語版 BART 要約タスクのファインチューニングの条件と定量的評価スコア

実験名	抽出前データセット数	抽出条件①データ文字数	抽出条件②文字数圧縮率	抽出後データセット数	ファインチューニング		スコア							
					epoch	model	BERT Score p	BERT Score r	BERT Score F1	BLUE	ROUGE-1	ROUGE-2	ROUGE-N	ROUGE-BE
BART1	wikiHow（27,877件）	0～1500文字	5～50%	19,740件	5	base	0.683	0.708	0.695	0.354	0.114	0.245	0.066	7.359
BART2	livedoor (101,559件)	0～1500文字	5～50%	84,635件	5	base	0.684	0.636	0.659	0.217	0.058	0.159	0.030	1.814
BART3	livedoor (101,559件)	0～1500文字	5～50%	84,635件	5	large	0.681	0.702	0.691	0.343	0.106	0.239	0.068	6.821
BART4	livedoor + wikihow (136,933件)	300～1000字	10～50%	56,734件	5	base	0.683	0.698	0.690	0.344	0.106	0.237	0.049	6.601
BART5	livedoor + wikihow (136,933件)	300～1500字	5～50%	81,177件	5	large	**0.692**	**0.720**	**0.706**	**0.386**	**0.144**	**0.275**	**0.082**	**9.671**

表 3. 各自動要約モデルの定量的評価スコア

要約手法	スコア							
	BERT Score p	BERT Score r	BERT Score F1	BLUE	ROUGE-1	ROUGE-2	ROUGE-N	ROUGE-BE
BART	0.687	0.718	0.702	0.368	0.126	0.258	0.082	8.225
T5	**0.728**	**0.738**	**0.733**	**0.438**	**0.185**	**0.328**	**0.091**	**14.351**
LexRank	0.665	0.694	0.679	0.344	0.096	0.211	0.045	4.71
AutoAbstractor	0.663	0.702	0.681	0.345	0.106	0.224	0.053	6.417

ここで，文字数圧縮率は，3.1 節で取得したデータセットから以下の式(1)により計算した．

$$文字数圧縮率 = \frac{要約後文字数}{要約前文字数} \times 100 \qquad (1)$$

上記の 5 つの条件を変化させることで最適なファインチューニング条件を探索した．探索したファインチューニング条件と定量的評価の結果の一部を表 2 に示す．表 2 より，wikiHow と livedoor を合わせたデータセットを要約前文字数 300~1500 文字，文字数圧縮率 5~50%で抽出し，BART large モデルで 5epoch 学習したときに最もスコアが高くなることが分かる．

4 他手法との比較

4.1 比較要約モデルの概要

4 節では，3.4 節において最もスコアの高い条件でファインチューニングした BART モデルと他の自動要約手法との比較を行う．比較の評価指標として，3.3 節における 4 種類の定量的評価とアンケート調査による定性的評価の 2 通り行った．比較する要約手法は，抽象型要約と抽出型要約のそれぞれについて検証した．他手法の抽象型要約モデルとして，本研究では，日本語の事前学習済みモデルが公開されている T5[12]との比較を行った．T5 はテキストを入力されるとテキストを出力するという統一的枠組みで様々な自然言語処理タスクを解く深層学習モデルである．また，他手法の抽出型要約モデルとして，LexRank[13]と AutoAbstractor[14]と比較を行った．LexRank は文章からグラフ構造を作り出して重要な文のランキングを作ることで要約する手法である．各文章を TF-IDF を用いて特徴ベクトルとして表現し、コサイン類似度を用いて文章間の類似度を計算し，多くの文章と類似度が高い場合、その文章は重要であると判断し，上位 3 文を抽出する．AutoAbstractor は，NLP ベースの要約手法で，入力文書全体の単語の出現頻度から文の重要度を計算する手法である． SimilarityFilter 機能を用いて，文章内にある文字列に対し類似性の尺度を使って計算し冗長な文章を削除したのち，重要度の高い文を要約として，上位 3 文を出力した．

4.2 定量的評価

表 3 に 4 種類の自動要約モデルの定量的評価スコアの結果を示す． T5 ＞BART ＞AutoAbstractor ＞ LexRank の順によいスコアとなった．また，抽象型要約の方が抽出型要約よりも良い結果となった．T5 が最もスコアがよい理由として，T5 の事前学習済みモデルは，Wikipedia の日本語ダンプデータ ，OSCAR の日本語コーパス，CC-100 の日本語コーパス合わせて約 100GB を用いているのに対して，日本語版 BART は約 1800 万分の日本語 Wikipedia の約 3GB を用いた事前学習済みモデルであるため，事前学習済みモデルの大きさが影響した可能性が高い．

4.3 定性的評価

金融文書を取得するため SharedResearch[15]から約 1000~1500 字の 3 つのアナリストレポートを引用し，BART，T5，AutoAbstractor の 3 つの手法に対して自動要約を実行した．

次に，生成された要約文に対して以下の 5 点の定性的評価項目を元にアンケートを作成した．

- 可読性：非文法的な繋がりが生じていないか．
- 了解性：意味をなさない文となっていないか．
- 忠実性：原文とは違う解釈の余地が生じていないか．
- 十分性：原文に含まれていた別の命題内容や、主題・陳述内容が欠落していないか．
- 非冗長性：要約文章に重複する内容がないか．

アンケートにおいては，回答者の答えやすさを考慮し，可読性と了解性については，「日本語が自然かどうか」，忠実性と十分性と非冗長性については，「要約文として適切かどうか」という 2 つの項目について調査した．

以下の図 2 は，「日本語が自然かどうか」について，図 3 は「要約文として適切かどうか」についてのアンケート調査結果の棒グラフである．アンケートの回答者数は 19 人で，3 つの手法それぞれで生成された要約文に対して，日本語が自然だと思う順番に順番を付ける形式で 3 つのアナリストレポート（Q1，Q2，Q3）に対して行った．グラフは，1 番目：3 点，2 番目：2 点，3 番目：1 点 として集計した棒グラフである．

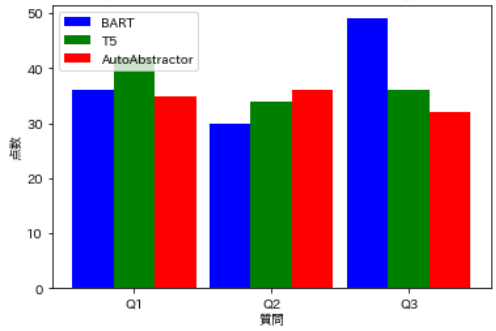

図 2. 定性的評価【文章の自然さ】棒グラフ

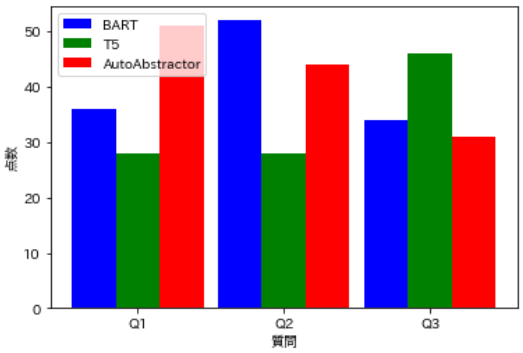

図 3. 定性的評価【文章の自然さ】棒グラフ

図 2 より，抽象型要約の方が文章の自然さは高くなる傾向があることがわかる．これは，抽象型要約のメリットである文章に違和感がなく，読みやすいという特性と一致している．

図 3 より，抽象型要約の方が文章の自然さは高くなる傾向があることがわかる．これは，抽象型要約のメリットである文章に違和感がなく，読みやすいという特性と一致している．

5 投資家向け支援システムの概要

抽象型文書要約を利用した以下の初心者向け投資先選択支援システムの開発に取り組む．

①利用者データの登録・金融文書の要約

利用者は投資にまつわる関心分野などをシステムに登録後，PDF ファイルや HTML 文書を入力することで，文書に含まれる文章を自動で抽出し，その要約文を得ることができる．

②金融文書の収取・投資先のレコメンデーション

TDnet などのオープンデータを活用し，金融文書を収集し，企業ごとにトピックモデルなどを用いた分析を行う．その結果を①の利用者に対して提供し，利用者の嗜好や関心を考慮した投資先を提案する協調フィルタリング手法を開発する．

このシステムにより，投資家は企業の経営状況だけでなく，より幅広い情報によって投資先を選定できる．

6 まとめ

本研究の貢献は以下の 3 つである．

1. Web スクレイピングによる要約用大規模データセットの収集を行った．
2. 収集したデータセットを様々な条件で抽出，ファインチューニングすることで日本語版 BART モデルの要約タスクにおける精度を定量的に評価した．
3. 金融文書に対して 4 種類の自動要約手法の定量的評価と定性的評価による比較し，金融文書の抽象型要約による初心者向け投資先選択支援システムを構築することでモデルの有用性を確認した．

今後の展望として，抽出型要約と抽象型要約を組み合わせたハイブリッド型の要約手法の提案や金融ドメインに特化したデータセットを用いたファインチューニングによる精度の比較を行いたい．また，5 章の投資家向け意思決定支援システムの構築とシステムの効果検証を行いたい．

参考文献

1. 平野 正徳, 坂地 泰紀, 松島 裕康, 和泉 潔, “金融文書のための別タスク学習による教師なし重要文判定”, 言語処理学会, pp.569‐572,(2020).

2. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer: BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension, In Proc. of ACL, pp.7871-7880,(2020).

3. Devlin, Jacob, et al. “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805,(2018).

4. Brown, T., et al. Language models are few-shot learners.Advances in Neural Information Processing Systems, 33, pp.1877-1901,(2020).

5. Satoru Katsumata, Mamoru Komachi. “Stronger baselines for grammatical error correction using a pretrained encoder-decoder model”. Asia-Pacific Chapter of the Association for Computational Linguistics, pp.827‐832,(2020).

6. 田中 佑, 村脇 有吾, 河原 大輔, 黒橋 禎夫, “日本語 Wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの改良”, 言語処理学会論文誌, 28 巻 4 号, pp.995-1033,(2021)

7. wikiHow, “信頼できるハウツーマニュアル”, https://www.wikihow.jp/, 最終アクセス日 2022/9/28

8. livedoor, “ライブドアニュース”, https://news.livedoor.com/, 最終閲覧日 2022/10/22

9. ThreeLineSummaryDataset, 3 行要約データセット, https://github.com/KodairaTomonori/ThreeLineSummaryDataset, 最終閲覧日 2022/10/22

10. Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In Proc. of EMNLP, pp.2292‐2297.Association for Computational Linguistics,(2015).

11. Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proc. of EMNLP, pp. 66‐71. Association for Computational Linguistics, (2018).

12. 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).

13. Gunes Erkan, Dragomir R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization” J. Artif. Int. Res., vol. 22, no. 1, pp. 457‐479, (2004).

14. 自動要約ライブラリ pysummarization, , https://code.accel-brain.com/Automatic-Summarization/, 最終閲覧日: 2022 年 1 月 2 日

15. SharedResearch, https://sharedresearch.jp/ja, 最終閲覧日: 2022 年 1 月 2 日