クエリ対応の事前要約を伴う大規模言語モデルによる企業事業概要生成

田村 光太郎
株式会社ユーザベース
koutarou.tamura@uzabase.com, k.tamura.phd@gmail.com

概要

企業の事業・財務の内容を把握することは、投資や経営の意思決定などビジネスにおいて、さまざまな場面で必要となる。一つの情報を多角的かつ高品質な情報として整理する必要があるが、実際には、多種多様なフォーマットで書かれた開示資料や公開情報などの読取りには、専門的な知識が必要となることもあり、人手で行うことに大きなコストがかかる。

本研究では、有価証券報告書や決算説明会の書き起こしなど、これらの企業情報を一定の形式に整理し、事業の概要を説明する文を生成することを試みる。その生成の方式として、多種多様なテキストの特徴や文章量を均すために、テキストを事前に要約し、RAG を行うことで、生成される概要文の品質を高めることを行った。

1 はじめに

企業活動に関する情報は、さまざまな場面で利用される。たとえば、市場での投資家が投資判断を行うことや、事業の見通しをたて、経営計画をたてることなどに主に利用される。これらは、日々のニュースや決算のタイミングで取得される開示資料などが利用されるが、これらは、上場企業に限った情報としても、およそ 4000 社分の情報があり、大量かつ高頻度に処理する必要がある。

これらの膨大な資料を使い、企業分析を実施すると、企業価値の判断など金融工学や財務会計のような方法を使うほかに、テキストベースのものについては情報抽出や要約などの定性情報を使って、情報を整理することがよく行われる。

従来では、多量のテキストデータから機械的に必要な情報を抽出し、整理・構造化を行うことが多く、ニュースや経済情報、市場で開示される企業情報から、いち早く情報を分類・抽出 [1, 2, 3, 4, 5, 6, 7] することが必要とされてきた。

一方で、テキストデータをさまざまな観点で要約し、情報を整理することも行われる。このアプローチでは、機械学習やデコーダ型の深層学習モデルを利用することが多いが、生成 AI による検索拡張生成（RAG：Retrieval-Augmented Generation）[8] を利用することが運用には適当と考えられている。RAG は、LLM への事前学習による知識の追加を行うのではなく、生成に必要な外部知識を入力としてあたえることで、知識を効率的に拡張する方法である。RAG を用いることで、情報の構造化にとどまらない、高度な要約文の生成やユーザーからのクエリに応じて生成結果を変えるインタラクティブな機能を持たせることが可能とされている。

本研究に関連するものとして、著者は、企業に関する複数の情報を外部情報とし、生成 AI を使って事前に要約を行った外部情報で RAG を行うことで事業概要を生成するシステムを提案した [9]。著者による先行研究に準じ、生成する概要文の数を増やし、提案手法による品質向上を議論する。

本章では、本研究の背景を述べた。第 2 章では、利用するデータの詳細を述べ、3 章では、提案手法の概要を解説する。4 章では提案手法を実際のデータに適用し、5 章でその結果を述べる。最終章で、本研究をまとめる。

2 データセット

有価証券報告書

本論文では、2023 年に開示された有価証券報告書を扱う。本研究では、決算数値とともに事業の現況や課題や展望がテキストとして記載されていることから、第２【事業の状況】で記載される「1。経営方針、経営環境及び対処すべき課題等」「3。経営者による財政状態、経営成績

及びキャッシュ・フローの状況の分析」の項目の文章を利用する。

決算説明会の書き起こしデータ

決算説明会での経営層からのプレゼンテーションの内容やアナリストからの質疑応答などのやりとりが文字起こしされているテキストデータである。上場企業の一部において、投資家向けの情報として公開される。ここでは、有価証券報告書と対応する 2023 年の通期の決算説明会の書き起こしデータを取得して利用する。

データの構成は、プレゼンテーション部分と質疑応答部分に分かれる。プレゼンテーション部分は、対応する決算期の説明となるため、有価証券報告書の内容と類似した内容である。一方で、質疑応答部分では、実際に役員とアナリストととの間の会話が質問（Q）と応答（A）に分類されたテキストとなっている。さらに、あいさつ、言葉のつまりのほか話者が複雑に入れ替わる場面が存在するため、レポート形式とは文体が異なる。

質疑は、テキストを順に追うことで、基本的に Q→A の対となるやりとりで構成される。しかし、場合によっては Q → A→A、Q → Q → A などで質疑のラベルが振られているものもある。ここでは、Q から A を挟んで、次の Q の手前までを 1 つの質問としてチャンク化したデータの整形を行っている。

3 提案手法

本研究における課題として、企業の事業概要把握に必要となるテキストデータは、企業によって文章量や文体がさまざまであり、生成結果における情報の濃淡や文体に影響を与えてしまうことにある。そのため、事前に外部情報となるテキストデータを一定の形式に要約し、テキストの量や文体を均一化する処理を行う方式を提案する。ただし、事前に要約する際に、最終的な要約に必要となる情報が欠落しないように、クエリに応じた事前要約を行うことで、外部情報の成形を行うものである。

具体的な提案方式の表式としては、下記のようになる。

$$p_{RAG}(y|x) = \sum_{z,z'} p_\eta(z|x) p_\theta(z'|x、d) p_\theta(y|x、z')$$

ここで、x はシステムに与えられるクエリ、z は外部情報源となるテキスト、y は最終的な生成物（ここでは事業概要文）である。この RAG システムは、クエリ x に応じて、検索モデル p_η で外部情報 z を選ぶ。そして、クエリ x に応じて選ばれた各外部情報 z を要約する。最後に、事前に要約された z' とクエリ x を使い、最終的な要約 z を得る。

具体的なクエリの内容やプロセスは、実験設定に記述する。

4 実験設定

提案手法の効果を検証するために、上場企業のうち決算説明会書き起こしデータを公開している 40 社を選定し、それぞれの企業の事業の「環境」「課題」「展望」「強み」の 4 観点について、事業概要文を生成することを試みる。事前要約、要約において利用する LLM は、gpt-4o を利用する。また、前述のとおり、有価証券報告書や決算説明会書き起こしデータを利用する。

我々の提案手法の概念図が図 1となり、具体的なプロセスとしては、下記の通りとなる。下記、X は企業名、T は「環境」「課題」「展望」「強み」から選ぶ。

- 企業 X と観点 T を定め、クエリ①としてシステムに与える。
- 企業 X に紐づくデータをデータベースから選出する。
- 上記データをクエリ②によって、観点 T に応じた事前要約でデータの均一化を施す。
- 事前要約したデータ群を使って、企業 X の観点 T における要約文の生成をクエリ③によって実施する。

提案手法におけるクエリ①においては、企業名 X や「環境」「課題」「展望」「強み」の各観点 T をクエリの一部として与えている。クエリ②、クエリ③は付録に記載する。

クエリ①における入力内容

X 社に関して、与えた各資料から事業の T の要約を完成させてください。

以上のように、企業名と観点を与えて入力したクエリに対し、ある企業 X の外部情報である有価証券報告書と決算説明会書き起こしデータ[1)]を選び、事

1) 本研究においては、企業に対して外部情報が有価証券報告書と決算説明会書き起こしデータと自明に紐づくため、提案手法における検索モデル p_η は、本研究においては、ルールベースで行う。

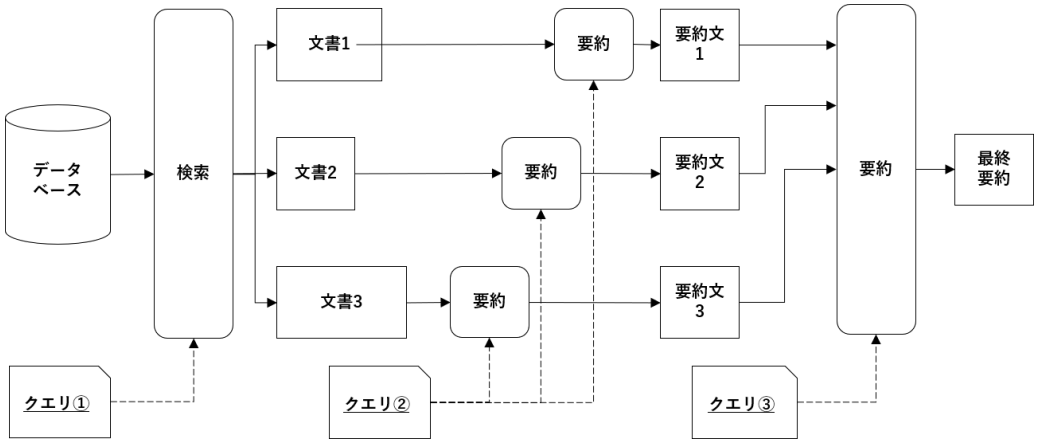

図 1: 提案手法におけるクエリの発生から生成までのプロセス。四角はデータ、角丸四角は LLM、角欠四角はクエリを表す。LLM はデータとクエリを受け取り、データを生成する。

前要約を行う。

特に、事前要約では、表 1 のようなトークン数となるようにデータを加工する。LLM で扱うテキストサイズが入力制限を超えるものについて、入力制限である 3 万トークンでデータを分割し、結果を統合する方式をとっている。

また、提案手法での比較のために、観点 T を指定して事前要約した外部情報での概要文生成（提案手法）と観点 T を指定しない事前要約（単純要約）での概要文生成を実施して比較する。与えるクエリは、付録における②' とする。生成文を数値的に比較することは難しいため、定性的な評価を行う。

5 結果

結果を文章の構造や妥当性としての評価と総論としての評価を行った。

5.1 文章の構造や妥当性

全体的に、単純要約では財務数値について言及することが多く、40 社のうち 4 観点とも財務数値を言及しているものは 29 社あり、財務的な話題に無理やり関連付けたような生成結果もうち 13 件ほど見られた。

提案手法や単純要約のどちらでも、言及している単語や表現は与えた入力文章の一部が転載・要約される傾向が高く、誤用などは見られなかった。しかし、企業のセグメント名（事業部名）の改変や、「AI ソフトウェア企業である○○」といった修飾表現が意図せずついてしまうケースが提案手法においても 6 件確認された。

全体として、単純要約や提案手法ともに会話文や口語の文章が混ざることなく生成がなされた。これは、事前の要約において外部情報の文体が統一されたことによるものが大きく、事前要約をすること自体に生成結果の文調を制御できることが示唆された。

一方で、観点を指定せずに要約させた単純要約において、有価証券報告書や決算説明会の書き起こしデータが財務数値の列挙や理念などを中心に要約されるケースが 40 社すべてで起き、クエリを指定しないで事前要約を行うと、事前要約時点で情報が欠落してしまうことが頻発することが分かった。これらのケースが単純要約での概要文の質を顕著にさげていた。

5.2 総論としての評価

前述のような細かい点での評価のほかに、文の形式・構造やコンテンツとしての評価を総合して、提案手法と単純要約の生成結果を比較したときに事業概要文としてどちらが妥当かという点で評価を行っている。表 2のように人手による生成文をチェックした結果として、提案手法が優位であることが見られた。

有価証券報告書や決算説明会書き起こしデータでは、「課題」「環境」に関する記述が多いため、これらの観点を適切に抽出できた提案手法が優位となった。しかしながら、提案手法のなかには、テキスト中で複数種の「課題」「環境」が言及された場合に、主要なものの抽出に失敗していたケースがあり、これらは単純要約が優位となっていた。そのほか、細部の記述が乏しく、情報として端的すぎる場合において、「どちらともいえない」と評価されるケース

表 1: 各データにおける事前要約する長さ

データ	項目	要約する長さ（トークン）
有価証券報告書	経営方針、経営環境及び対処すべき課題等	2000
	経営者による財政状態、経営成績及びキャッシュ・フローの状況の分析	2000
決算書き起こし	プレゼンテーション	2000
	質疑応答データの各チャンク	500

表 2: 提案手法の評価

観点	提案手法優位	どちらともいえない	単純要約優位
課題	26	7	7
環境	27	7	6
展望	18	11	11
強み	17	12	11

が多かった。

一方で、「展望」「強み」に関する記述は、元データにおいて企業ごとに記載の有無や量に差がある。結果的に、当該観点の概要文を生成すること自体が難しいものであると考えられるが、それでも一定程度の提案手法の優位性が見て取れた。

6 まとめ

本研究において、企業概要文の生成を行う際の課題に焦点を当てた。フォーマットやサイズが異なる複数のデータソースを用いる RAG において、入力データの均一化を図ることと、記述したい観点を整理する事前要約のプロセスを提案した。観点を指定するか否かに関わらず、事前要約自体に生成結果の文体を整える効果があることが示唆された。これは過去研究 [10, 11] と整合する。さらに、提案手法として観点を指定して事前要約することで、要約時に重要な情報が欠落してしまうことを防ぎ、最終的な生成結果の質を挙げられることが示唆された。特に、著者の先行研究と比較しデータ量を増やした実験を行い、改めて提案手法の優位を確認できた。

今後の課題として、事前要約による扱うトークン数の増大への対処や生成結果の定量的な評価が挙げられる。前者においては、多段的な要約をおこなうことによるトークン数の増加があり、テキストデータのチャンク化やそれに基づく部分的な検索の実施が考えられる。後者では、重要単語を固有表現抽出や関係抽出で抜き出すことで、確認する部分をより簡略化したり、どのテキストを参照しているかの透明性を確保するために元データとの類似箇所の判定も運用上必要とされる。自動評価技術の検討に対し、これまで当該分野で行われてきた情報抽出技術が利用されることも考えられる。

さいごに、本研究は LLM を利用したデータの統合と高品質な生成にむけた事前要約の知見を示したことで、提案手法が RAG を使った生成の一手法として検討の余地を与えたものである。さらなる改善とその応用の可能性のために、手法の検証を行っていきたい。

参考文献

[1] 奥田裕樹, 髙橋寛治: ニュース記事からの企業キーワード抽出, 言語処理学会第 26 回年次大会

[2] 橋本航, 笛木正雄, 黒木裕鷹: 日本語固有表現抽出における *BERT-MRC* の検討, 言語処理学会第 28 回年次大会

[3] 高橋寛治, 甫立健悟, 奥田裕樹: ニュース記事中の組織名の曖昧性解消, 人工知能学会全国大会論文集 (2023)

[4] 田村光太郎, 北内啓, 高山温: 固有表現抽出によるニューステキスト内の企業名抽出, 人工知能学会全国大会論文集 (2023).

[5] 田村光太郎: 固有表現抽出タスクにおける文章のランダム連結によるデータ拡張, 2023-IFAT-153(3), (2023)

[6] 田村光太郎: 疑似ニュース生成による固有表現抽出タスクのデータ拡張, 第 19 回 Web インテリジェンスとインタラクション研究会, (2023)

[7] 澤田悠治, 安井雄一郎, 大内啓樹, 渡辺太郎, 石井昌之, 石原祥太郎, 山田剛, 進藤裕之, 日経企業 *ID* リンキングのための類似度ベース *EL* システムの構築と分析, 言語処理学会第 30 回年次大会

[8] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv. /abs/2005.11401, (2020)

[9] 田村光太郎: クエリに対応した事前の要約を伴う大規模言語モデルによる企業事業概要生成, 情報処理学会デジタルプラクティス, Vol.6 No.1, (2025)

[10] Optimizing Retrieval-Augmented Generation with Advanced Chunking Techniques: A Comparative Study, https://antematter.io/blogs/optimizing-rag-advanced-chunking-techniques-study

[11] Matouš Eibich, Shivay Nagpal, Alexander Fred-Ojala, *ARAGOG: Advanced RAG Output Grading*, arXiv:2404.01037

付録

生成した概要文の例

著者らの先行研究 [9] より引用する。

事業課題として生成された結果：単純要約

日本ハムの経営方針は社会貢献を目指し、従業員の幸せと生き甲斐を追求するとともに、高品質の食品を提供することにより、消費者の健康な生活に貢献することを目指しています。その一環として、2023 年の中期経営計画では、売上高 1 兆 2、600 億円、事業利益 380 億円などの目標を（中略）加工事業部が若干のマイナスを見込んでいる一方で、食肉、海外、ボールパーク・その他の事業部がプラスを見込んでいると述べています。これらの結果を受けて、企業は今後の事業戦略として、売上高を通期で 1 兆 3、000 億円、事業利益を 440 億円と見込んでいます。

事業課題として生成された結果：提案手法

日本ハムは、社会的な変化や環境課題に対応するために、事業の強化と構造改革、持続可能性の追求、新規事業の育成を進めています。その一方で、原材料やエネルギー価格の高騰、消費者の節約志向による低価格帯へのシフトなどにより、経営環境は厳しさを増しており、（中略）価格高騰や売り場の縮小による在庫管理の問題や、労働コストの上昇などが挙げられます。経営課題の解決に向けては、適切な在庫管理やマーケティング戦略、新商品開発、労働力の確保などが重要となります。また、構造改革や事業の再編も視野に入れており、これらの取り組みにより、事業の持続的な成長を目指しています。

提案システムで利用するクエリ

クエリの番号は、図 1に対応する。また、クエリ中のトークン数 L は、表 1のものとなる。

クエリ②における入力内容

以下の内容で、X 社の事業の T を要約するために必要な内容を、L トークン以内の文章に要約してください。（以下、各テキストデータを挿入）

クエリ③における入力内容

この X 社に関する下記の文章を使って、X 社の事業の T についての要約を完成させてください。（以下、事前要約された各テキストを連結して挿入）

クエリ②’ における入力内容

以下の内容を、L トークン以内の文章に要約してください。（以下、各テキストデータを挿入）