有価証券報告書の活用による事業セグメント関連語の拡張

伊藤友貴 [1] 平松賢士 [2]
[1] 三井物産株式会社 [2] 株式会社アイフィスジャパン
[1]Tomok.ito@mitsui.com [2]kenji.hiramatsu@ifis.co.jp

概要

各企業はIR（Investor Relations）活動として，決算短信や有価証券報告書等を通し，機関投資家や個人投資家等へ企業の情報を開示する．IR 活動において，自社の発信内容に対する外部からの評価の把握は重要な事項である．特に事業セグメント単位での評価の把握は有用だと考えられる．このような背景のもと，本研究では，金融文書からの「事業セグメント言及文抽出手法」の開発を目指す．セグメント情報の抽出に関する既存アプローチとして，事業セグメント名の検索をベースとするアプローチがある．一方，本アプローチでは関係会社の記載やサービス名・取扱商品のみが書かれた文の抽出には対応できないという課題がある．そこで，本研究では大規模言語モデル及び有価証券報告書内の事業セグメント説明記載を利用することで検索単語を拡張し，既存アプローチを改善することを提案する．さらに，本取組みの発展として，アナリストレポートのセグメント別センチメント分析や決算短信へのアナリスト反応の生成を試みる[1)]．

1 はじめに

各企業はIR（Investor Relations）活動として，決算短信や有価証券報告書，統合報告書等の開示を通し，機関投資家や個人投資家等へ企業の情報を開示する．このIR活動は，透明性の確保，企業と投資家との間の信頼関係の構築，市場の期待値の管理，危機管理，競争優位の獲得，等の上で重要な活動の一つである [1]．特に 2023 年 1 月 31 日，金融庁が「企業内容等の開示に関する内閣府令」等の改正を公表した．結果，各企業はESG や SDGs 等の非財務情報の開示も求められており、各企業の IR 活動は益々重要なものになると想定される．ここで，各企業の IR 担当者にとって「自社 IR が機関投資家にどう捉えられるかを把握できること」は重要な課題の一つであると考えられる．特に，株主還元への施策や各事業セグメントの説明に関する自社発信内容に対する外部専門家の評価を各項目毎に把握できるとは各企業の IR 担当者にとって有用であると期待される．

このような背景のもと，本研究では IR 文書やアナリストレポート等の金融文書からの「事業セグメント言及文抽出手法」の開発を目指す．特に，新規上場会社や組織改組等により事業セグメントが変更された場合における活用を考慮し，教師なしでの抽出手法開発を目指す．ここで，アナリストレポートとは，証券アナリストが企業の経営状態や収益力などを分析・調査し，まとめたレポートのことであり，その記載内容が株式リターンや各企業の業績の増減に関連することが報告されている [10, 9]．例えば，アナリストレポートと決算短信，双方から事業セグメントに関して言及している文を抽出し，セグメント単位で紐づけることで，自社発信内容に対する外部の反応を把握することが可能となる．

「教師なし」での「事業セグメント言及文抽出」に関する取組みの一つに「セグメント名の検索」をベースにしたアプローチ [6, 7] があり，その有効性も検証されている．一方，本アプローチでは関連会社やサービス名・商材等のみを記載する文の抽出には対応できないという課題がある．

そこで，本研究では，有価証券報告書上の事業セグメント説明記載から各事業セグメントに関連する単語を取得し，検索に利用する単語を良質な形で拡張することで「関連会社やサービス名・商材等にも対応可能」にすることを提案する．関連語の取得には，自然文への対応のため大規模言語モデル（LLM）を活用する．実データを用いた検証の結果，提案手法を用いることで「事業セグメント名」をベースとする既存アプローチに比べ高い性能で「事業セグメント言及文」を抽出できることを実証できた．

さらに発展として，アナリストレポートの「セグメント別センチメント分析」及び「決算短信からの

1） 本予稿上の記載内容は著者個人の見解であり所属組織の意向には関係ありません．

アナリスト反応（セグメント別）の生成」を試みた．

2 関連研究

IR 情報の分析に関する取組みとして，例えば [4] では決算説明会議事録から抽出される感情極性と株式リターンに関する分析を行い，質疑応答セクションの極性とイベント (決算説明会) 後の収益率の相関を分析した．また [5] では有価証券報告書や統合報告書から各企業の「環境活動」に関する情報を抽出の上，各活動の自動評価，及び活動の改善案を自動生成を試みている．

IR 文書からのセグメント情報抽出に関連する取組みや手法の提案も既にいくつか報告されている．[6] では「事業セグメント名」の検索をベースに業績要因文・業績結果文の抽出を行っている．また，[8] では，事前に用意された学習データに基づく，教師あり学習をベースとする抽出手法を提案している．これらの手法は有効である一方，[8] のような教師あり学習を前提とする手法では訓練データを構築する必要があり，その利用範囲が限定される．例えば新規上場会社や組織改組等により事業セグメメントが変更された場合に，[8] のアプローチでの対応は難しい．また，[6] のような「事業セグメント名」の検索をベースとするアプローチに関しては 1 節にて説明した課題がある．

3 事業セグメント関連語の拡張

本節では有価証券報告書内の事業セグメントに関する説明記載に対し LLM を活用することで事業セグメント文抽出時の検索単語を良質に拡張することを提案する．本手法は以下 2 Step から構成される．

Step 1: 有価証券報告書からの「事業セグメント説明記載」の抽出（3.1 節）

Step 2: 事業セグメント関連語の抽出（3.2 節）

IR 文書やアナリストレポート等の文書から各事業セグメントに関する記述を文単位で抽出する際に，事業セグメント名のみを検索の対象とする既存のアプローチ [6] では，サービス名、及び関係会社名のみが記載される文には対応できない．上記のように検索の範囲を拡げることで，このような文にも対応できることが期待される．

3.1 有価証券報告書からの事業セグメント説明記載の抽出

本 Step ではまず「事業セグメント説明記載」を有価証券報告書（PDF) から抽出する．有価証券報告書では多くの場合，図 1 のように表形式で事業セグメントのサービス名や関係会社名が記載される．図 1 の例[2)]では各サービス名や会社名が句点で区切られており処理が容易だが，処理が難しい形で記載される例も多くある．本ステップではまずこの表をルールベースで抽出する．具体的には【事業の内容】と記載されるページ及びその次のページから pymupdf[3)] を用いて表を抽出する．その後「セグメント」という文字が記載の列，または左端の列からセグメント名を検索し，各行の文字列を全て抽出する．

3.2 事業セグメント関連語の抽出

Step 1 の出力結果を用いて，以下の形式のプロンプトを LLM に入力することで，事業セグメント関連語として「サービス名」「取り扱い商品名」及び「関係会社名」を出力する．

質問 次の取扱商品またはサービスの内容及び関係会社に関する説明から取扱商品名、サービス名、または関係会社名を全て抽出し、',' 区切りて出力してください。

説明 事業セグメント説明記載（Step 1 出力結果)

出力された文字列をカンマ (,) または句点（、）で区切り，事業セグメント関連語を抽出する．文書内の各文に関してこの事業セグメント関連語が含まれた文を抽出し，事業セグメント言及文を抽出する．

4 事業セグメント言及文抽出

本節では，前節にて提案した「事業セグメント関連語の拡張」の手法を利用することで，「事業セグメント言及文抽出」の性能を向上できるか否かを実データを用いて検証する．

4.1 検証データ

検証データには「アナリストレポート」及び「決算短信（IR)」の二つのデータセットを用いた．

2) https://www.mitsui.com/jp/ja/ir/library/securities/___icsFiles/afieldfile/2020/06/19/ja__101yuho.pdf
3) https://pymupdf.readthedocs.io/ja/latest/

当社グループの連結決算対象会社の総数は506社であり、その内訳は連結子会社が海外209社、国内74社、持分法適用会社が海外178社、国内45社となっています。

セグメント	取扱商品またはサービスの内容	主要な子会社	主要な持分法適用会社
鉄鋼製品	インフラ鋼材、自動車部品、エネルギー鋼材 他	三井物産スチール、Regency Steel Asia、Game Changer Holdings、EURO-MIT STAAL、Bangkok Coil Center	日鉄物産、GRI Renewable Industries、Shanghai Bao-Mit Steel Distribution、Gestamp North America、Gestamp Holding Mexico、Gestamp Brasil Industria De Autopecas、Gestamp Holding Argentina、GESTAMP 2020、SIAM YAMATO STEEL、GEG (Holdings)
金属資源	鉄鉱石、石炭、銅、ニッケル、アルミニウム、製鋼原料・環境リサイクル 他	Mitsui-Itochu Iron、Mitsui Iron Ore Development、Mitsui Iron Ore Corporation、Mitsui & Co. Iron Ore Exploration & Mining、Oriente Copper Netherlands、Japan Collahuasi Resources、三井物産カッパーインベストメント、三井物産メタルズ、Mitsui & Co. Mineral Resources Development (Asia)、Mitsui Coal Holdings、Mitsui & Co. Mozambique Coal Investment、Mitsui & Co. Mozambique Coal Finance、Mitsui & Co. Nacala Infrastructure Investment、Mitsui & Co. Nacala Infrastructure Finance	Inner Mongolia Erdos Electric Power & Metallurgical、日本アマゾンアルミニウム、BHP Billiton Mitsui Coal

図 1　事業セグメントに関する記載例．三井物産株式会社第 101 期有価証券報告書

4.1.1　アナリストレポート

本データセットには，総合商社セクター 5 社（8001 伊藤忠，8002 丸紅，8031，三井物産，8053 住友商事，8058 三菱商事）発行の各 2020 年度 4Q 決算短信に対し，証券会社 3 社から発行されたアナリストレポート計 44 レポート（各商社 4〜5 レポートずつ）に記載の文が含まれる．各文へは各事業セグメントの内容を「含む」または「含まない」のラベルが付与され，本データセットは 291 の「含む」ラベルと 4,285 の「含まない」ラベルから構成される．

4.1.2　決算短信（IR）

本データセットには総合商社セクター 5 社（8001 伊藤忠，8002 丸紅，8031，三井物産，8053 住友商事，8058 三菱商事）の 2020 年度 4Q 決算短信（連結）に記載される文が含まれる．各文へは各事業セグメントの内容を「含む」または「含まない」のラベルが付与され，本データセットは 501 の「含む」ラベルと 7,278 の「含まない」ラベルから構成される．

4.2　ベースライン

性能検証のため，以下の手法と性能比較をした．

GPT 3.5:　gpt-3.5-turbo[4)] によるゼロショット推論．

ベースライン:　事業セグメント名のみを検索に利用する手法．

評価指標には「含む」ラベルの抽出に対する F1 値を用いた．提案手法では 2020 年度 3Q 以前に発行かつセグメントの説明が記載された有価証券報告書の中で最新のものを利用した．

4.3　検証結果

検証結果は表 1 及び表 2 の通りである．提案手法の有効性が実証された．また，抽出結果を見ると，ベースライン手法では三菱商事株式会社の関係会社であるローソンに関する記載が抽出に失敗したものの，提案手法では成功している等の事象が見られた．提案手法による検索単語の拡張が，性能改善につながったものと思われる．

5　応用タスクへの活用

更に「アナリストレポートのセグメント別センチメント分析」や「アナリスト反応の生成」等の応用タスクへの提案手法の活用可能性を検証した．

5.1　セグメント別センチメント分析

本検証では，まず 3 節にて提案の手法と同様に「事業セグメント関連語を含む段落」を抽出する．

4）　https://platform.openai.com/docs/models/gpt-3-5

表 1　セグメント言及文抽出結果 (アナリストレポート)

	Precision	Recall	F1
GPT 3.5	0.894	0.320	0.471
ベースライン	0.927	0.395	0.554
提案手法	0.840	0.760	0.798

表 2　セグメント言及文抽出結果 (決算短信)

	Precision	Recall	F1
GPT 3.5	0.876	0.204	0.331
ベースライン	0.874	0.331	0.481
提案手法	0.821	0.459	0.589

その後，抽出した段落に対し，「各事業セグメント」に関するセンチメントが「ポジティブ (+1)」「ネガティブ (-1)」「ニュートラル (0)」のいずれであるかを LLM を用いて推定する．その後，その結果を平均することで各ラベルを付与する．尚，LLM の推論は GPT 3.5 (ゼロショット) で実施した．また，「事業セグメント関連語を含まない段落」については「記載なし」と判定する．

5.1.1　データセット

検証データには 4 節で利用したアナリストレポートから構築された，データセットを用いた．本データセットには各段落に対し，以下のように各セグメント単位でのセンチメントが付与されている．

{" 金属資源": " ポジティブ", " 鉄鋼製品": " ネガティブ", ..., " 化学品": " 記載なし"}

ラベルの種類は「ポジティブ」「ネガティブ」「ニュートラル」「記載なし」の 4 種であり，ラベル数の内訳は 160, 44, 33, 1,422 である．

5.1.2　ベースライン

性能検証のため，以下の手法と性能比較をした．評価指標には F1 値 （マクロ平均）を利用した．

GPT 3.5:　GPT 3.5 を活用したゼロショット推論.

ベースライン:　段落抽出時にセグメント名のみ使用する手法.

5.1.3　検証結果

結果は，表 3 の通りである．提案手法を用いると，ベースライン手法から F 1 値が向上することを確認できた．

表 3　セグメント別センチメント分析検証結果

	Precision	Recall	F1
GPT 3.5	0.428	0.386	0.392
ベースライン	0.761	0.510	0.585
提案手法	0.637	0.624	0.628

5.2　決算短信からのアナリスト反応生成

本取組みの更なる延長として「自社発信内容に対する外部専門家の各事業セグメント反応の生成」を試みた．まず，提案手法を活用し，決算短信及び対応するアナリストレポート双方から，各事業セグメントに関する記載内容を抽出した．その後，この「決算短信上の記載」と「対応するアナリストレポート上の記載」のセグメントレベルでの組に関するデータセットを用いて「自社発信内容に対する外部専門家の各事業セグメントの反応」を生成した．生成には，本文中学習（事例数 5) を利用した few-shot 推論を利用した．学習サンプルは 2019 年度 4Q に発行された総合商社 5 社の決算短信，及びそれに対応するアナリストレポート 10 本から，提案手法を用いて生成された「各セグメント単位での決算短信上での記載（入力)」と「各セグメント単位でのアナリストレポート上での記載（出力)」の組を用いた．出力結果やその定性評価等は当日紹介する．

6　結論

本研究では、有価証券報告書の事業内容に関する説明記載に関して LLM を活用し適切に変換した上，検索時に利用する事業セグメント関連語を拡張することで「金融文書からの事業セグメント情報の抽出手法」を改善することを提案した．また，実データを用いた検証により提案手法の有効性を実証した．さらに，発展として応用タスクである「アナリストレポートのセグメント別センチメント分析」及び「決算短信記載内容からのアナリストのセグメント別コメント生成」への提案手法の活用を試みると共に，活用時における提案手法の有効性を実証した．

一方，今回，実施した検証は総合商社セクターのみに留まり，本手法の他セクターへの有効性に関する検証は今後の課題となる．また，各手法の性能向上，統合報告書やサステナビリティレポート等への拡張や抽出項目の拡張等も今後の課題となる．

参考文献

[1] 東京証券取引所,「コーポレートガバナンス・コード〜会社の持続的な成長と中長期的な企業価値の向上のために〜」, 東京証券取引所, 2018.

[2] Kei Nakagawa, Shingo Sashida, Ryozo Kitajima, and Hiroyuki Sakai. What do good integrated reports tell us?: An empirical study of japanese companies using text-mining. In 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 516–521. IEEE, 2020.

[3] 片山 喬博, 特集「非財務情報開示の潮流」, PwC's View, 第 42 号, 2023. https://www.pwc.com/jp/ja/knowledge/prmagazine/pwcs-view/202302/42-01.html.

[4] 黒木裕鷹, 真鍋友則, 指田晋吾, 中川慧. 決算説明会テキストデータの感情極性と株式リターンの分析. 人工知能学会第二種研究会資料, Vol. 2022, No. FIN-029, pp. 47–53, 2022.

[5] 児玉実優, 酒井 浩之, 永並健吾, 高野海斗, 中川 慧, 企業における環境活動の改善案の自動生成, 人工知能学会第二種研究会資料, Vol. 2023, No.FIN-031, pp. 75–80, 2023.

[6] 高野海斗, 酒井浩之, 北島 良三, 有価証券報告書からの事業セグメント付与された業績要因文・業績結果文の抽出, 人工知能学会論文誌, Vol. 34, 2019.

[7] 伊藤友貴, 小林暁雄, 関根聡, 決算短信からの事業セグメント情報抽出, 言語処理学会第 24 回年次大会, 2018.

[8] Tomoki Ito, Hiroki Sakaji, Kiyoshi Izumi, Segment Information Extraction from Financial Annual Reports Using Neural Network," Annual Conference of the Japanese Society for Artificial Intelligence, pp.215-226, Niigata, Japan, June, 2019.

[9] 北島良三, 酒井浩之, 上村龍太郎, 坂地泰紀, 平松賢士, 栗田昌孝, アナリストレポートと企業業績の関係解析 (第一報), 人工知能学会第 22 回金融情報学研究会, pp.53-56, 2019.

[10] 平松賢士, 酒井浩之, 坂地泰紀, 極性付与されたアナリストレポートと株式リターンとの関連性, 人工知能学会第 20 回金融情報学研究会, pp.54-60, 2018.

[11] 平松賢士, 三輪宏太郎, 酒井浩之, 坂地泰紀, アナリストレポートのトーンの情報価値, 証券アナリストジャーナル, Vol.59, No.2, pp.86-97, 2021.