アナリストレポートにおける キーワード関連文の抽出と景況感推移観測への応用

Extraction of Keyword Related Sentences in Analyst Reports and Its Application to Observation of Business Confidence

高山 将丈[1]　小澤 誠一[1,2*]　廣瀬 勇秀[3]
飯塚 正昭[3]　渡辺 一男[3]　逸見 龍太[3]
Shota TAKAYAMA[1]　Seiichi OZAWA[1,2]　Takehide HIROSE[3]
Masaaki IIZUKA[3]　Kazuo WATANABE[3]　Ryuta HENMI[3]

[1] 神戸大学大学院 工学研究科
[1] Graduate School of Engineering, Kobe University
[2] 神戸大学 数理・データサイエンスセンター
[2] Center for Mathematical and Data Sciences, Kobe University
[3] 三井住友 DS アセットマネジメント株式会社
[3] Sumitomo Mitsui DS Asset Management Company, Limited

Abstract: Analysts in investment trust management companies survey business achievements of target companies and are supposed to summarize them in the form of an analyst report. A fund manager reviews the reports and decides which companies to invest based on the reports as well as various economic indices and information. However, when searching for potential investment targets, a fund manager is generally required to read a large number of analyst reports and other related documents, Obviously, this work is not an easy task even for a skilled manager. In this work, we propose an intelligent system that retrieves meaningful sentences related to a specific query such as 'performance' and automatically evaluates a market trend in order to mitigate their work loads. From a total of 37,398 analyst reports and interview records, we obtained word embedding vectors using Word2Vec, and related sentences addressing company's financial soundness were retrieved based on the similarity to a query. In our experiments, for the word 'achievement', we retrieved 395 sentences out of 2,182 sentences that were 2.67 times larger than those when an exact search was applied. On the other hand, a trend of market sentiment obtained from a keyword such as 'profit' did not have high correlation against actual market indices.

1　はじめに

投資信託を請け負う投資信託運用会社（運用会社）は，投資によって利益が得られる優良企業を見出し，その投資額を適切に設定することが求められる．このため，運用会社に所属するアナリスト及びファンドマネージャーは，日常的に投資候補企業の調査を行い，投資先の選定を行っている．具体的には，まずアナリストが投資候補企業の調査を行い，その調査結果を速報的にアナリスト往訪記録という文書にまとめ，十分に情報収集ができたタイミングで有価証券報告書や株価，同業他社の情報なども分析して，正式な報告書であるアナリストレポートを発行する．このアナリストレポートをもとに，ファンドマネージャーは投資候補企業への投資判断を行っている．この投資先決定フローにおいて，ファンドマネージャーは，アナリストがまとめた多数のレポートを精読しなければならず，これには多大な労力と時間を要する．そこで，この業務をサポートするシステムの開発が求められている．

このようなシステムとして，深層学習を用いたデータ駆動型の文書解析手法が注目されており，アナリストレポートの文書表現から景況感を読み取る試みが行われている．小林ら [1] は，アナリストレポートからアナリスト予想の根拠を含む文を自動抽出できること

*連絡先：神戸大学 数理・データサイエンスセンター
〒 657-8501 神戸市灘区六甲台町 1-1
E-mail: ozawasei@kobe-u.ac.jp

を示した．また，得られた文に対して極性付与を行い，75.13%の精度で極性を推定できることを示した．小林らの手法では，「手がかり表現」と呼ばれるキー単語を人手で入力して，アナリスト予想の根拠となる文の特徴抽出が試みられている．そして，得られた特徴量に対して 12 層の全結合ニューラルネットワークを利用し，景況感判定を行っている．この手法では，根拠情報を含む文しか出力できないことや事前に定義した「手がかり表現」を入力する必要があり，分析者がアナリストレポートの特徴を熟知していることが前提となる．また，多数の全結合層で構成されたニューラルネットワークを使用しているため，勾配消失や学習の発散，過学習などが懸念される．

本研究では，分析者の専門知識を前提としなくても自動的に手がかり表現の関係性が学習され，指定したキー単語に依存しない関連文抽出が可能となる深層学習モデルのアプローチを採用する．ニューラルネットの構造を制約することで，全結合ニューラルネットワークで問題となる勾配消失等の問題を緩和し，文書の分類タスクで著しい成果を挙げている「BERT[2]」を用いたシステムを提案する．

2　提案システム

本論文では，アナリストレポートにおいて，蓄積されたアナリストレポートを用いて，入力された視点に対する景況感の推移を確認できるシステムを提案する．

本システムの概略図を図 1 に示す．本提案システムは大きく分けて 2 つのモジュールで構成されており，それぞれ図 1 の，課題 2 で示されている部分である．1 つ目は Word2Vec[3] を用いて入力したキーワードにおける関連文をアナリストレポートから抽出するモジュールである．2 つ目は自然言語処理の各種タスクにおいて目覚ましい成果を挙げている BERT を用いたモジュールで，1 つ目のモジュールで得られた関連文をもとにキーワードにおける景況感推移を観測するものである．

2.1　キーワード関連文抽出モジュール

例えば，「業績」という観点で景況感追跡を行いたい場合，「業績」という単語を持つ文を抽出するだけでは不十分であることは想像に難くない．「業績」に関して述べるために「決算」や「収益」など，「業績」に近い意味を持つ単語を用いている可能性も低くないからである．

本モジュールの概略図を図 2 に示す．これは「業績」だけでなく「決算」や「収益」などの単語を含んだ文の抽出を行うことを目的とした．本モジュールに入力した「業績」などのキーワードと近い意味を持つ「決算」や「収益」などの単語を，Word2Vec を用いて抽出し，それらの単語を含む文を関連文としてアナリストレポートから抽出することによって実現する．

まずアナリストレポート及びアナリスト往訪記録をコーパスとして学習させた Word2Vec を用いて，入力したキーワードと近い意味を持つ単語 (関連語) を抽出する．上述の例においては，キーワードが「業績」，関連語が「決算」や「収益」となる．そして，キーワードと関連語を含む文を抽出すること本モジュールの出力とする．

このようにして得られたキーワード関連文は，次の景況感推移観測モジュールに入力される．

2.2　景況感推移観測モジュール

本モジュールは，自然言語処理の各種タスクにおいて目覚ましい成果を挙げている BERT を用いたモジュールで，キーワード関連文抽出モジュールで得られた関連文をもとにキーワードにおける景況感推移を観測するものである．概略図を図 3 に示す．

まず Wikipedia コーパスを用いて BERT の事前学習を行う．その後景況感ラベルの付与されたアナリスト文書を用いて BERT の Fine-Tuning を行い，文章から景況感を推定できるように学習を行う．これによって得られた学習済み BERT モデルを用いて，キーワード関連文抽出モジュールから入力された文ごとに景況感判定を行い，年度内で景況感を平均する．これによってキーワードにおける年度ごとの景況感が抽出可能となる．

3　実験設定

3.1　検証に利用したデータ

本提案システムは特定企業におけるキーワード別の景況感を追跡するものである．本検証では東証一部上場企業 1 社について検証を行った．この企業を企業 AB と定義する．存在したアナリストレポートは，2002 年から 2016 年に発行されたもので，総文書数は 76 文書，総文数は 1764 文であった．アナリストレポートは docx ファイルで発行されていたため，Python モジュールである docx2txt[1] を利用し，文字列の抽出を行った．

3.2　キーワード関連文抽出モジュール

三井住友 DS アセットマネジメント株式会社から提供を受けたアナリストレポート及びアナリスト往訪記

[1] docx2txt · PyPI, https://pypi.org/project/docx2txt/

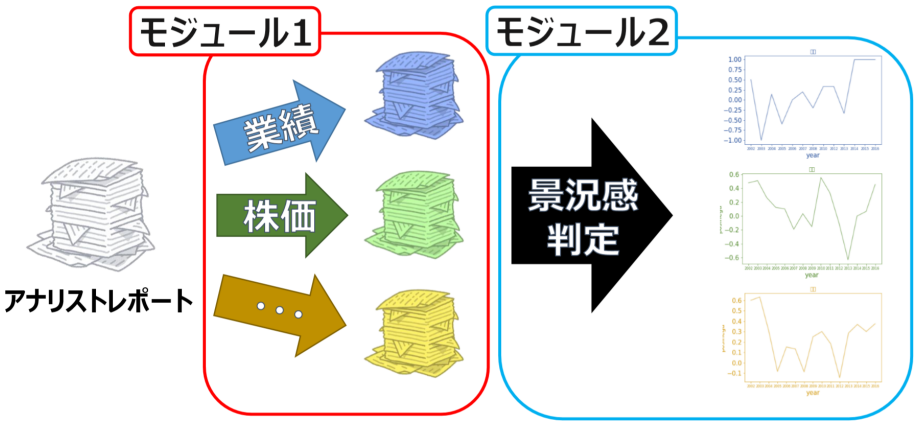

図 1: 提案手法の概略図

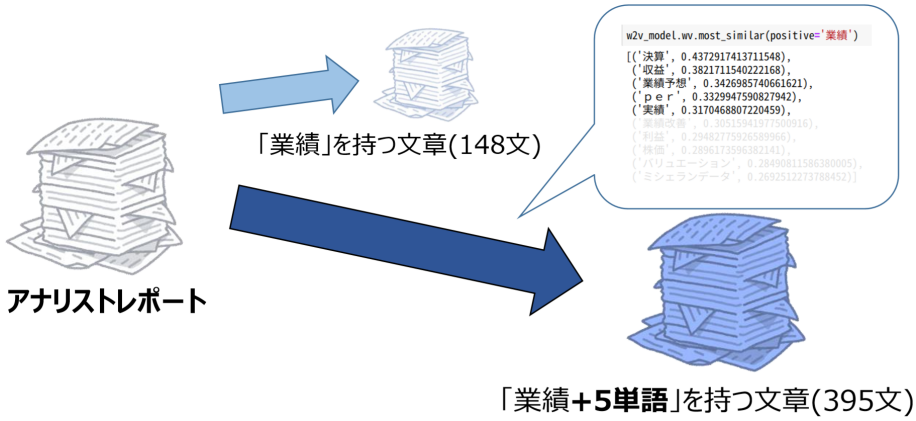

図 2: キーワード関連文抽出モジュールの概略図

表 1: アナリスト文書の内訳

文書の種類	ラベル	数量
アナリスト往訪記録	ポジティブ	1,145
	ネガティブ	788
	ニュートラル	2,160
	-	9,580
アナリストレポート	-	21,892

表 2: Word2Vec のパラメーター

パラメーター名	値
ウィンドウサイズ	2
単語ベクトルの次元数	500
min_count	1
イテレーション数	20

録をコーパスとして，Word2Vec の学習を行った．利用した文書データの内訳及び Word2Vec の学習のパラメーターをそれぞれ表 1 及び表 2 に示す．

また学習前の処理として，これらの文書に適切な前処理を行った後，活用する単語の原形への変換，数値データの正規化，記号の削除の処理を行った．分かち書きには Python モジュールである MeCab を利用し，その辞書として NEologd[4] を利用した．

学習後の Word2Vec を用いてキーワードに対する関連語抽出するが，本検証では Word2Vec の出力する単語ベクトルのうち，キーワードのベクトルにコサイン類似度の近い上位 5 つの単語を関連語と定義した．

3.3 景況感推移観測モジュール

本検証ではウェブ上で一般公開されている事前学習済み BERT モデルを利用した[2]．これは Wikipedia コーパスを用いて学習させたもので，SentencePiece[5] で

[2]BERT with SentencePiece を日本語 Wikipedia で学習してモデルを公開しました，https://yoheikikuta.github.io/bert-japanese/

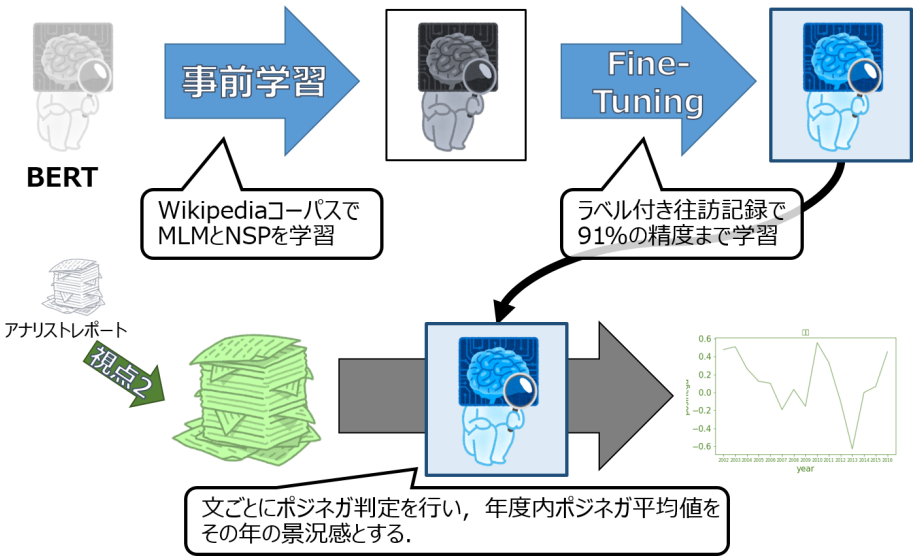

図 3: 景況感推移観測モジュールの概略図

表 3: BERT のパラメーター

パラメーター名	値
最大入力トークン数	398
バッチサイズ	2
BERT 内部の隠れ層の次元数	768
学習率	1e-5
ドロップアウト率	0.1
BERT 内部の隠れ層の活性化関数	GELU[6]
Attention[7] のヘッド数	12
Encoder[8] の層数	12
語彙数	32,000

表 4: 「業績」の関連語

単語	コサイン類似度
収益	0.487
決算	0.479
利益	0.479
ﾊﾞﾘｭｴｰｼｮﾝ	0.460
業績予想	0.453
上期決算	0.437
株価	0.405
受注販売	0.385
受注	0.382
単独決算	0.367

トークン化を行ったものである．SentencePiece の辞書は 32,000 の語彙数で学習されたものである．

本検証における BERT のパラメーターは表 3 のとおりである．

Wikipedia コーパスを用いて事前学習を行った後，キーワード関連文抽出モジュールで得られた文ごとに景況感判定を行うため，Fine-Tuning を行った．BERT の最終層に全結合層を結合し，景況感となる 2 値分類を文ごとに行う．アナリスト往訪記録のうちラベルデータの付与された 1,933 文書を用い，テストデータにおいて 91%程度の分類精度が可能となるまで学習を行った．

4 性能評価

4.1 キーワード関連文抽出モジュール

まず，本モジュールで得られた関連語を表 4 及び表 5 に示す．

直感的な違和感を覚えることはほとんどないと言える．特に，「業績」を入力することによって得られた「バリュエーション」は投資判断において一つの重要な指標を示す単語であり，この単語が抽出されていることは注目に値する．「利益」においても「営業利益」・「経常利益」・「純利益」などの単語が抽出され，類似語を抽出するという目標は達成されたと言える．

次に，「利益」を入力することで抽出された関連語から得られた関連文例を図 4 に示す．

2 文目では，「利益」の単語が含まれていないにも関わ

- 0 年度営業利益実績推定値と [G 製品] 事業の有利子負債削減額が事前ｺﾝｾﾝｻｽを上回った点はｻﾌﾟﾗｲｽﾞ。
- ｸﾞﾙｰﾌﾟﾛｰﾝを圧縮することで、[H 製品] 事業の収益力は更に向上が見込めるとしている。
- 会社の優先順位は、期間損益 (pl) が赤字になってしまうよりもﾌﾘｰｷｬｯｼｭﾌﾛｰをﾌﾟﾗｽにすることで、販売状況がさらに悪化すれば、在庫調整のｽﾋﾟｰﾄﾞはさらに強化され、[I 事業所] の効率は大きく低下するであろう。

図 4: 「利益」の関連文例

表 5: 「利益」の関連語

単語	コサイン類似度
収益	0.629
損益	0.534
営業利益	0.521
経常利益	0.509
純利益	0.494
粗利益	0.485
業績	0.479
最終損益	0.467
税引後利益	0.464
売上	0.460

らず，利益に関係が深いと言える文が抽出されている．しかし 3 文目では，「利益」よりもフリーキャッシュフローを重視しているという内容で，利益に関係が深い文とは言えない．このような文も一部抽出されており，除外したり影響を小さくしたりするシステムを開発する必要があると考えられる．

4.2 景況感推移観測モジュール

まず，キーワード関連文抽出モジュールで得られた各文に対して BERT で景況感判定を行った結果を図 5 に示す．

上から 2 つの文に関しては得られた景況感に関して大きな違和感は無く，正しく推定が行われていることがわかる．しかし 3 番目の文のように，1 文内に景況感が混在している文も存在し，人間によっても景況感を付与することは難しい．また 4 番目の文では「売上未達」のネガティブ情報と，「収益性の向上」のポジティブ情報が両方含まれている．文末にあるポジティブ情報を強調した文であると考えられるが，システムの推定結果はネガティブ情報に注目していると考えられる．5 文目は景況感情報が含まれていない情報である．また，例示はしていないが，過去に発行したアナリストレポートから引用されている場合もあり，これらの文章を景況感判定に用いてしまうと，常に同じ結果を出力してしまうため，景況感のバイアスとなってしまい，変化点の検出が難しくなる．

このように，抽出された関連文に景況感と直接関係のない文やバイアスとなる文章，1 文内に景況感が混在する文などが複数含まれており，これらの文を除外したり，係り受け解析などの手法を用いてさらに細かい文書単位で景況感判定を行う必要があると考えられる．また，4 文目のような文を正しく処理するため，文末に重要な情報を配置する日本語の特性を踏まえたシステムを開発する必要があると考えられる．

次に，景況感を追跡した結果を図 6 に示す．企業 A の純利益の実データと，本システムのキーワードとして「利益」を入力した結果を比較した．純利益の実データは有価証券報告書から取得したものである．

システムから得られた景況感推移と対応する実データの推移に明確な相関を発見することはできなかった．

抽出された文には，会社発表やアナリストの想定と実績との差に関しての記述が含まれている．減益であったとしても想定通りならポジティブな文体で書かれており，増益であったとしても想定を下回るようならばネガティブな文体で書かれている．システムから得られた景況感が大きく振れたタイミングで得られた文を解析することで，これらの差を生み出した要因の抽出やアナリストの深層心理や勘違いを可視化することが可能なシステムとなる可能性がある．

5 結論

本論文では，投資信託運用会社の業務をサポートするため，過去のアナリストレポートを解析し，任意の企業・視点に対する景況感を追跡するシステムを提案した．具体的には，与えたキーワードに基づいてアナリストレポートから関連文を抽出するキーワード関連文抽出モジュールと抽出された関連文の景況感判定を

推定結果:ネガティブ　当社では現在の状況から会社計画の達成はやはり困難、会社計画比 0 億円強の営業利益下ぶれを想定する。

推定結果:ポジティブ　① [A 製品] の拡大・ [B 製品] のｼｪｱｱｯﾌﾟ・ｺｽﾄ削減による収益力強化が続く。

推定結果:ポジティブ　部門利益予想は [C 製品] 事業を 0 億円引き上げ、[D 製品] 部門を 0 億円引き下げた。

推定結果:ネガティブ　会社計画比で、減収増益を予想する主な要因は、[C 製品] 事業において、[E 地域] の [C 製品] の販売減速による売上未達と、[F 国] での値上げの浸透による収益性の向上にある。

推定結果:ポジティブ　【今年の株価維持策】…0 月の説明会は既に案内が届いている。

図 5: BERT における各文の景況感推定結果例

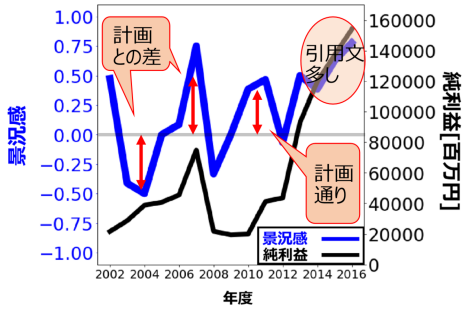

図 6: 企業 A における景況感抽出結果と実データの比較

行って，時期ごとにその推移を観測する景況感推移観測モジュールを開発した．これらモジュールによって，大量に蓄積されたアナリスト往訪記録やアナリストレポートから関心の高い記述を見つけることが容易になり，ファンドマネージャーやアナリストの業務をサポートする当初の目標を一定程度実現できた．

しかし，関連文の抽出精度がキーワードによっては，あまり良くないという課題も残されている．また，得られた景況感と実際の経済指標との明確な相関を確認できず，この点も課題として残されている．一方で，本システムにおける関連度抽出の精度がさらに向上すれば，アナリストの深層心理を可視化できる可能性もあり，更に検証を行っていく必要がある．

参考文献

[1] 小林和正，酒井浩之，坂地泰紀，平松賢士．アナリストレポートからのアナリスト予想根拠情報の抽出と極性付与．第 *19* 回 人工知能学会 金融情報学研究会資料, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[4] Sato Toshinori. Neologism dictionary based on the language resources on the web for unidic-mecab, 2015.

[5] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on EMNLP: System Demonstrations*, pages 66–71. ACL, 2018.

[6] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*, 2016.

[7] Cicero Nogueira dos Santos Mo Yu Bing Xiang Bowen Zhou Yoshua Bengio Zhouhan Lin, Minwei Feng. A structured self-attentive sentence embedding. *Technical report*, 2017.

[8] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on EMNLP*, pages 1724–1734. Association for Computational Linguistics, 2014.