債券市場における金融極性辞書の自動構築とその拡張

Automatic Construction and Expansion of Financial Sentiment Lexicons on Bond Market

今井 康太[1]　酒井 浩之[1]　永並 健吾[1]　稲垣 真太郎[2]
Kota Imai[1], Hiroyuki Sakai[1], Kengo Enami[1], Shintaro Inagaki[2]

[1]成蹊大学
[1]Seikei University
[2]みずほ証券株式会社
[2]Mizuho Securities Co., Ltd.

Abstract: 本研究では，債券市場を対象とした金融極性辞書の自動構築手法を提案する．これまでに株式市場に特化した金融極性辞書の研究は行われているが，債券市場には対応できていない．そこで，本研究では債券市場を対象とし，語の前後関係まで拡張した文字列を収録した金融極性辞書の自動構築手法を考案した．さらに，自動構築した金融極性辞書の一部を学習データとし，文中の形態素にタグを付与するモデルを活用することによって新たな表現を獲得し，金融極性辞書の拡張を行う．

1. はじめに

投資家は様々なデータを踏まえて投資判断を行うが，そのデータは膨大で多岐にわたるため，近年は膨大な金融情報を分析して投資判断を支援する技術が注目されている．その一例が「利上げ」のような金融に関連する語を収録した金融極性辞書である．金融の分野では，企業の発行する決算短信，有価証券報告書，統合報告書など，日々，膨大な金融に関する文書を読み分析する必要がある．そのような文書中には市場における極性が付与できる語が含まれている．例えば「利上げ」は日銀が利上げを実施すれば株価は下がるため，株式市場にとってはネガティブな語である．市場分析では金融に関する文書中に含まれるそのような語の極性から市場予測をする等があるが，金融に関連する語は専門的な語も含めて数多く，分析には専門的な知識を要する．金融極性辞書とは，そのような語に対して極性が付与された辞書であり，市場分析の支援に有用な情報である．

金融極性辞書について，これまでに株式に特化した「ネガティブ・ポジティブ」付与の研究が行われている[1][2][3]．しかし，その極性辞書を他の金融市場に当てはめると極性が異なる場合が多い．これは同じ表現であっても金融市場によって極性が異なる場合が存在するからである．例えば，「物価上昇」は株式市場においてはポジティブと判断するが，債券市場においてはネガティブと判断する必要がある．なぜなら一般的には物価が上昇すれば景気が好調であり株価も上昇するが，その場合，株式市場に資金が流れ，リスクオンに伴い安全資産でもある債券の人気が下がるからである．

そこで，本研究では特に債券市場を対象として，複数の形態素で構成された表現に対して景気要因における極性，債券要因における極性をそれぞれ付与することで，上記の問題を解決することを試みる．

本研究では，更なる表現の獲得を行うために，文を構成する形態素ごとにタグ付けした学習データを用意することにより，新たな表現の獲得をする．このような手法は，主に固有表現抽出のタスクで用いられており，既存の研究で有効性が示されている[4][5]．しかし，形態素ごとにタグが付与されたデータが必要であるため，新たなタスクに取り組む際に

は，特に学習データの作成に大きな労力がかかる．そこで，本研究では，最初に自動的に作成される金融極性辞書の表現を用いることで，タグ付与のために使用する学習データを自動で生成することにより，人手で学習データを作成せずに新たな表現の獲得を行い，金融極性辞書の拡張を行う．

2. 関連研究

本研究の特徴として，文を構成する形態素にタグを付与することにより，新たな表現を獲得しているという点が挙げられるが，同様の手法を用いた研究がいくつか存在する．Li らは，BERT ベースモデルで単語ごとに極性分析を行う手法を提案した[4]．また，Arkhipov らは，多言語固有表現抽出タスクにおいて，BERT ベースモデルに対して，対象となる言語で再学習することにより，精度を向上させる手法を提案した[5]．これらの研究は，人手で作成された学習データを用いているが，本研究では，最初に自動的に作成される表現を用いて形態素ごとにタグ付与をすることで，学習データを自動生成している点が異なる．

3. BERT モデル

BERT[6]は，Google によって開発された自然言語処理モデルであり，ラベルなしデータを用いた事前学習と，比較的少量のラベル付きデータを用いたファインチューニングの 2 段階で学習を行うことによって，高い精度と汎用性を実現している．

本研究で使用する BERT モデルは東北大学の乾研究室で公開されている事前学習済み日本語 BERT モデル[1]に対して，タスクごとにファインチューニングをしたものを使用する．

3.1. 2 値分類 BERT モデル

入力シーケンスの先頭に[CLS]というトークンを追加することによって，[CLS]トークンのベクトル値により文などのシーケンスデータを 2 値分類することができる．主に文の極性分析に用いられる．本研究では，コーパスに含まれる文の分類と，文への極性付与をする際に 2 値分類 BERT モデルを使用する．図 1 にモデル図を示す．

3.2. タグ付け BERT モデル

シーケンスに含まれる各トークンのベクトル値を使用して，トークンごとにタグを付与することができる．タグを単語の品詞にすることによって表現の抽出をするなど，主に固有表現抽出に用いられる．本研究では，コーパスに含まれる表現の抽出をする際にタグ付け BERT モデルを使用する．図 2 にモデル図を示す．

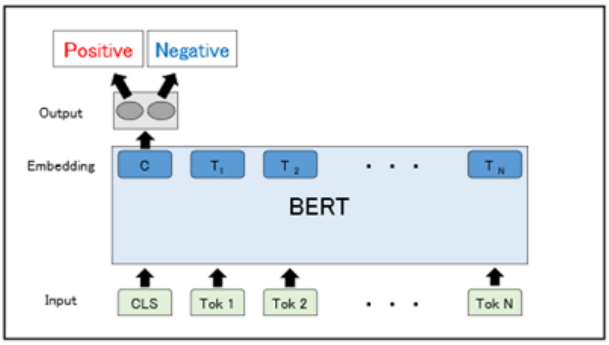

図 1 2 値分類 BERT モデル

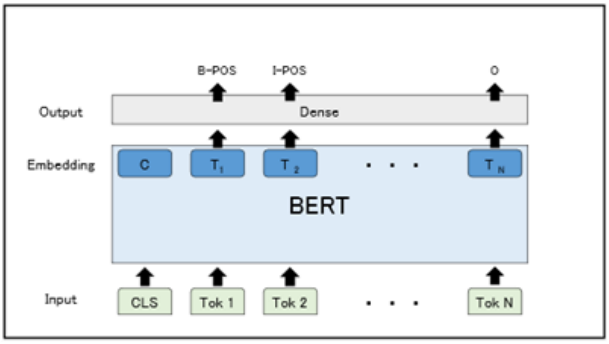

図 2 タグ付け BERT モデル

4. 提案手法の概要

本研究では，「みずほマーケットレポート」[7]を用いて，債券市場における表現にまで拡張した文字列

[1] https://github.com/cl-tohoku/bert-japanese

での金融極性辞書の自動構築を行う手法を提案する．提案手法の概要を以下に示す．

Step 1: みずほマーケットレポートのうち，債券市場について記載されている文書と景気動向について記載されている文書（以降，両方のデータを合わせたものを「債券文書データ」と定義）を抽出する．

Step 2: 債券文書データを用いて，債券市場における金融極性辞書に収録するのに適した語の候補（以降，「金融市場特徴語候補」と定義）を抽出する．

Step 3: 酒井らの手法[8]を用いて獲得した手がかり表現と金融市場特徴語候補をもとに，特徴的な語を表現にまで拡張した文字列（以降，「金融市場特徴表現」と定義）を獲得する．

Step 4: 獲得した金融市場特徴表現に極性の付与を行い，債券市場における金融極性辞書を構築する．

Step 5: Step 4 で構築した金融極性辞書に含まれる表現を用いて学習データを自動的に作成し，タグ付け BERT モデルによって債券文書データから新たな極性付与済みの金融市場特徴表現を獲得する．

5. 金融市場特徴表現の獲得

本研究では，表現の複雑な極性に対応することが目的であるため，語の前後関係まで拡張した文字列を表現として獲得したい．そのため，まず金融極性辞書に収録するのに適した語である金融市場特徴語候補と，その語を抽出するための手がかりとなる表現を債券文書データから自動的に抽出する．そして，抽出した金融市場特徴語候補と手がかりとなる表現を結合した文字列を，金融市場特徴表現として獲得する．

5.1. 金融市場特徴語候補の抽出

債券市場特有の表現を獲得するために，金融市場特徴語候補を債券文書データから抽出する．手法の概要を以下に示す．

Step 1: みずほマーケットレポートから債券文書データを抽出する．

Step 2: 抽出した債券文書データをもとに Word2vec のモデルを生成する．

Step 3: 生成した Word2vec のモデルに獲得した共通頻出表現を入力し，モデルから出力された語を金融市場特徴語候補として抽出する．

5.2. 金融市場特徴表現の獲得手法

抽出した手がかり表現と金融市場特徴語候補をもとに，金融市場特徴表現を自動獲得する．獲得手法の概要を以下に示す．

Step 1: 5.1 節で抽出した金融市場特徴語候補と手がかり表現を結合し，表現にまで拡張した文字列を生成する．

Step 2: 生成した文字列が債券文書データに含まれている場合のみ，その文字列を金融市場特徴表現として獲得する．

以下に獲得した金融市場特徴表現の例を示す．

> 後押しが強かった，出荷が不振，輸出が拡大した，成長期待が弱い，倒産リスクが拡大，生産性が低下

6. 金融市場特徴表現への極性付与

「はじめに」において述べたように，債券市場を対象とした金融市場特徴表現の極性は，表現の対象が景気であるか債券であるかによって異なる．すなわち，同じ表現であっても，景気を対象としている場合と債券を対象としている場合では異なる極性を付与する必要がある．そのため，表現に対する極性付与を景気要因と債券要因で別々に行う．そこで，債券文書データを景気要因に関する文と債券要因に関する文に分類し，それぞれの文集合を作成する．そして，2 つの文集合のそれぞれの文に極性を付与し，極性付与された文を用いて，含まれている金融市場特徴表現に極性を付与する．

6.1. 債券文書データにおける文の分類手法

債券文書データに含まれる文を，景気要因に関す

る文と債券要因に関する文に分類する手法について述べる．分類手法の概要を以下に示す．

Step 1: 2 値分類 BERT モデルにより，債券文書データにおける文を，景気要因または債券要因に関する文と，それ以外の文に分類する．

Step 2: 2 値分類 BERT モデルにより，Step 1 で景気要因または債券要因に関する文として分類された文集合を，景気要因に関する文と債券要因に関する文に分類する．

6.2. 金融市場特徴表現への極性付与手法

5 章で述べた手法により獲得した金融市場特徴表現に対して，景気要因と債券要因のそれぞれを対象とした極性の付与を行う．極性付与手法の概要を以下に示す．

Step 1: 抽出した手がかり表現に対して，手がかり表現のみで極性が明らかなものに対して人手で極性の付与を行う．

Step 2: 景気要因に関する文の集合から，文中に金融市場特徴表現が含まれている文のみ抽出する．(以降，「テストデータ」と定義)

Step 3: テストデータの中から Step 1 で極性を付与した手がかり表現が含まれている文に対してその極性を付与し，学習データを自動作成する．

Step 4: 2 値分類 BERT モデルを用いて，テストデータに極性の付与を行う．

Step 5: テストデータに含まれる金融市場特徴表現に対して，その金融市場特徴表現が含まれているテストデータである文に付与された極性の尤度をもとに極性の付与を行う．

Step 6: Step 1 から Step 5 までを債券要因でも同様に行う．

6.3. 金融市場特徴表現への極性付与結果

BERT によって極性付与されたテストデータ各文のネガティブ，ポジティブの尤度とその文に含まれる金融市場特徴表現をもとに，金融市場特徴表現に対して極性の付与を行う．極性付与された金融市場特徴表現の例を表 1 に示す．

表 1 極性が付与された金融市場特徴表現の例

極性 / 要因	ポジティブ	ネガティブ
景気	経済活動が弱まる 景気後退が鮮明	景気楽観論が台頭 好材料ではある
債券	国債買い入れが拡大 債券買いが進む	債券安が起きる 株式相場が上昇

7. 金融極性辞書の拡張

前章までで債券市場における金融極性辞書の構築を行ったが，使用したコーパスである債券文書データにはまだ獲得できていない表現が多く含まれている．更なる表現の獲得を行うため，構築した金融極性辞書の表現を学習データとして活用し，タグ付け BERT モデルを用いることによって辞書を拡張する．手法としては固有表現抽出のタスクに使われるタグ付け BERT モデルの適用であるが，そのために必要な学習データは，前章までで獲得した金融市場特徴表現を用いることで自動生成する．それにより，人手で作成することは困難な量の学習データを用いることができる．金融市場特徴表現拡張手法の概要を以下に示す．

Step 1: 景気要因の文集合の各文を形態素解析し，トークンごとにタグを付与できる状態にする．

Step 2: 景気要因の文集合から，極性の付与をした金融市場特徴表現を含む文のみを抽出する．

Step 3: 各金融市場特徴表現に対応する極性を用いて，Step 2 で抽出した文に対してトークンごとにタグ付けを行う．

Step 4: タグ付けされた文とそれに対応するタグのリストを学習データとして，BERT で学習を行う．

Step 5: BERT モデルを用いて，景気要因の文集合の各文に対してトークンごとにタグを付与する．

Step 6: 付与されたタグに基づいて表現の抽出を行い，新たに出現した表現のみを獲得する．

Step 7: Step 1 から Step 6 までを債券要因でも同様に行う．

7.1. トークンへのタグ付け

学習データを作成するために，トークンごとに分割したテキストデータに対して，トークンごとにタグを付与する必要がある．本研究では 7 つのタグを設定することにより，特徴的な表現と共にその表現の極性も獲得する．トークンに付与されるタグは，前章までで獲得した極性が付与された金融市場特徴表現と一致しているかどうかによって決まる．表 2 にタグの詳細を示す．表 2 に示したタグとその意味に基づき,各トークンに対応するタグの付与を行い，学習データを生成する．

表 2 タグの詳細

クラス番号	タグ	意味
0	O	表現に含まれない
1	B-POS	ポジティブな表現の始まり
2	I-POS	ポジティブな表現の継続
3	E-POS	ポジティブな表現の終わり
4	B-NEG	ネガティブな表現の始まり
5	I-NEG	ネガティブな表現の継続
6	E-NEG	ネガティブな表現の終わり

7.2. 新たな金融市場特徴表現の獲得

作成した学習データを 3 章で説明したタグ付け BERT モデルで学習し，その学習済みモデルで各文に対してトークンごとにタグの付与を行う.その後，付与されたタグに従って，新たに出現した表現のみを獲得する.表 3 に新たに獲得した表現の例を示す.

表 3 新たに獲得した表現の例

極性 / 要因	ポジティブ	ネガティブ
景気	停滞感が漂う 動きが弱い	株高が生じる 景気が下支え
債券	コア国債が堅調 5 年債入札が堅調	増加ペースが鈍化 関心が少ない

8. 評価

本研究における評価は，獲得した金融市場特徴表現の精度と表現に付与された極性の精度の 2 点について行う．

8.1. 金融市場特徴表現の評価

獲得した表現が金融市場特徴表現として適切であるかの評価を行う．評価方法の概要を以下に示す．

Step 1: 2 値分類 BERT モデルで獲得した表現のリストから，景気要因と債券要因のそれぞれで表現をランダムに 100 個ずつ抽出する．

Step 2: ランダムに抽出した各表現において，その表現が金融市場特有の表現であるかを人手で判定を行い，提案手法の精度を求める．

Step 3: タグ付け BERT モデルで獲得した表現においても，同様の流れで手法の精度を求める．

表 4 に金融市場特徴表現の評価結果を示す．また，各手法で獲得できた表現の数を併記する．なお，タグ付け BERT で獲得した表現は 2 値分類 BERT には含まれていない．従って，表現の異なり数は，2 値分類 BERT で獲得された 22,466 表現とタグ付け BERT で獲得された 20,758 表現の合計 43,224 表現となる．

表 4 金融市場特徴表現の評価結果

手法 / 要因		景気	債券	全体
2 値分類 BERT	精度(%)	91.0	89.0	90.0
	表現数(個)	9,036	13,430	22,466
タグ付け BERT	精度(%)	76.0	82.0	79.0
	表現数(個)	5,845	14,913	20,758

8.2. 金融市場特徴表現における極性の評価

金融市場特徴表現に付与された極性が適切であるかの評価を行う．評価方法の概要を以下に示す．

Step 1: 2 値分類 BERT モデルで獲得した表現のリストから，景気要因と債券要因のそれぞれで表現をランダムに 50 個ずつ抽出する．

Step 2: ランダムに抽出した各表現において，付与されている極性が適切であるかを人手で判定を行い，提案手法の精度を求める．

Step 3: タグ付け BERT モデルで獲得した表現においても，同様の流れで手法の精度を求める．

表 5 に金融市場特徴表現における極性の評価結果を示す.

表 5 金融市場特徴表現における極性付与の評価結果

手法＼要因		景気	債券	全体
2 値分類 BERT	精度(%)	80.0	68.0	74.0
タグ付け BERT		68.0	80.0	74.0

9. 考察

タグ付け BERT モデルによる表現獲得の精度について考察する．評価結果から，2 値分類 BERT で表現を獲得した時と比べて，多くの新しい表現が獲得できているが，精度が落ちていることが分かる．誤って獲得された表現として，例えば「矢印が示す」のような金融市場特有の表現ではないものがいくつか存在した．この問題の原因として，本研究ではタグ付け BERT モデルの学習データを自動で作成している点が挙げられる．データを自動生成することによって人手では難しい量の学習データを作成できるというメリットはあるが，精度への影響も考えなければならない．したがって，自動で辞書を作成した後に，人手で辞書に含まれている表現を選別する手法が良いと考える．

BERT モデルの構造について考察する．本研究で使用した BERT モデルは，最終層に全結合層を採用した．関連研究においては，全結合層を使用した場合と CRF を使用した場合との比較が行われており，結果から CRF の有効性が示されている[4][5]．これは，CRF がタグ間の依存性の学習をすることができるからだと考えられる．本研究では，タグの並びが不自然な場合は抽出をしないという処理を行ったが，CRF を採用することにより，更なる精度の向上と獲得できる表現数の増加が期待できる．

10. まとめ

本研究では，金融テキストを用いて，債券市場における表現にまで拡張した文字列での金融極性辞書を自動構築するための手法を提案した. 具体的には，みずほマーケットレポートを用いて，債券市場における金融市場特徴語候補を抽出し，手がかり表現と金融市場特徴語候補をもとに，金融市場特徴表現を獲得した．その後，獲得した金融市場特徴表現に極性の付与を行った．極性の付与は景気要因と債券要因で分けて行い，精度の向上を試みた．また，獲得した金融市場特徴表現を学習データとして，文のトークンへのタグ付けを行うモデルを活用することによって，新たな表現の獲得に成功した．

今後の課題として，現在の手法では文のトークンへのタグ付けを行う際にタグの並びを考慮できていないため, CRF を用いたモデルの改善を行うことで，精度の向上が期待できると考える．

参考文献

[1] 伊藤諒, 和泉潔, 須田真太郎: “ネットワークの表現学習による金融専門極性辞書の構築,” 2017 年度人工知能学会全国大会(第 31 回), 2017.

[2] 五島圭一, 高橋大志: “株式価格情報を用いた金融極性辞書の作成,” 自然言語処理, Vol.24, No.5, pp.547-577, 2017.

[3] 関和広, 柴本昌彦: “銘柄固有の金融極性辞書の構築,” 第 18 回金融情報学研究会, 2017.

[4] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam: “Exploiting BERT for End-to-End Aspect-based Sentiment Analysis,” in Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pp. 34–41, 2019.

[5] Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin: “Tuning multilingual transformers for named entity recognition on slavic languages,” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), p. 89–93, 2019.

[6] 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.

[7] みずほ証券株式会社: “みずほマーケットレポート,” みずほ証券株式会社 金融市場調査部レポート, 2000-2020.

[8] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀: “企業の決算短信 PDF からの業績要因の抽出,” 人工知能学会論文誌, Vol.30, No.1, pp.172-182, 2015.