

債券市場における金融極性辞書の自動構築

Automatic Construction of Financial Sentiment Lexicons on Bond

今井 康太¹ 酒井 浩之¹ 高野 海斗¹ 北島 良三²
末廣 徹³ 稲垣 真太郎³ 木村 柚里³

Kota Imai¹, Hiroyuki Sakai¹, Kaito Takano¹, Ryoza Kitajima²
Toru Suehiro³, Shintaro Inagaki³, Yuri Kimura³

¹成蹊大学

¹Seikei University

²東京工芸大学

²Tokyo Polytechnic University

³みずほ証券株式会社

³Mizuho Securities Co., Ltd.

Abstract: 本研究では、金融テキストを用いて、債券市場における金融極性辞書を構築するための表現獲得、および極性付与手法を提案する。獲得する語句は、各金融市場において特徴的な語やそれらを表現にまで拡張した文字列である。例えば、「増加」という語について、その語のみでは極性を決定できず、「消費が増加」や「リスクが増加」のように語の前後関係まで広く捉えることで極性が付与できる。我々はこれまで株式を対象として表現獲得、および、極性付与のための手法[1]を開発しており、良好な精度を得た。しかし、債券を対象として表現獲得、および、極性付与を行う場合、高い精度を達成できなかった。債券における表現への極性付与の精度が低い原因は、債券市場における極性付与が複雑であるという点が挙げられる。そこで、本研究では特に債券市場を対象として、表現に対して景気要因における極性、債券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。

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[2][3][4]。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。また、既存の金融極性辞書には単語（あるいは複合語）が収録されており、例えば、「増加」という語について、一つの極性が付与されている。しかし、「消費が増加」や「コストが増加」のように、語の前後関係まで広く捉えてみると、必ずしもその語に付与されている極性が正しいとは限らないことがわかる。

我々はこれまで株式を対象として表現獲得、および、極性付与のための手法[1]を開発しており、良好な精度を得た。しかし、債券を対象として表現獲得、および、極性付与を行う場合、高い精度を達成できなかった。債券における表現への極性付与の精度が低い原因は、債券市場における極性付与が複雑であるという点が挙げられ

る。例えば、景気が好調である場合は、一般的には資金需要が増加し、それに伴って金利が上昇して債券相場は下落する。そのため、例えば「米経済が回復する」のような景気についての表現では、株式ではポジティブであるが債券市場においてはネガティブの極性を付与しなければならない。そこで、本研究では特に債券市場を対象として、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。

これまでに金融極性辞書についての研究は行われている[2][3][4]が、株式に特化した辞書が多く、辞書に収録されているのは単語(あるいは複合語)単位である。本研究では、債券市場に着目している点と、語の前後関係まで拡張した表現を自動的に獲得する点が異なる。

2. 提案手法の概要

本研究では、日経 QUICK 月次調査と日本経済新聞の記事を用いて、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: 日経 QUICK 月次調査と日本経済新聞の記事を用いて、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 2: 酒井らの手法[5]を用いて獲得した手がかり表現と Step 1 で抽出した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 3: 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。

3. 金融市場特徴語候補の抽出

3.1 金融市場特徴語候補抽出の概要

Step 1: 日経 QUICK 月次調査に記載されている債券市場担当者からのコメント文(以降、「コメント文」と定義)を抽出し、コメント文と日本経済新聞記事をもとに Doc2vec のモデルを生成する。

Step 2: 生成した Doc2vec のモデルを用いて、債券市場

に関する記事を日本経済新聞記事から抽出し、コメント文と合わせた文章データ(以降、「拡張文章データ」と定義)を作成する。

Step 3: 酒井らの手法[5]を用いて、拡張文書データから共通頻出表現・手がかり表現を獲得する。

Step 4: 拡張文書データをもとに、Word2vec のモデルを生成する。

Step 5: 生成した Word2vec のモデルに獲得した共通頻出表現・手がかり表現を入力し、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。

3.2 Doc2vec のモデル生成と類似記事抽出

日経 QUICK 月次調査に記載のコメント文を抽出する。しかし、抽出したコメント文だけでは金融市場特徴語を獲得するデータ量として不足しているため、それらのテキストデータをもとに Doc2vec を用いて、日本経済新聞記事より債券市場に関する記事を抽出する。Doc2vec のモデルを生成し、日経 QUICK 月次調査に記載のコメント文集合と日経新聞記事との類似度を求め、上位 50,000 記事に関連記事としてコメント文と合わせて拡張文書データとする。

3.3 共通頻出表現・手がかり表現獲得

酒井らの手法[5]を用いて、金融市場特徴語候補を抽出するための手がかりとなる表現を拡張文書データから自動的に獲得する。共通頻出表現・手がかり表現を獲得する手法について以下に述べる。なお、具体的な獲得手法については、文献[5]を参照されたい。

Step 1: 少数の手がかり表現(今回は「が好調」「が不振」を用いる)を手で与え、それに係る節を取得する。

Step 2: 取得した節の集合から、その中で共通して頻繁に出現する表現(「売り上げ」など)を共通頻出表現として抽出する。

Step 3: 共通頻出表現に係る節を取得し、その中から新たな手がかり表現を抽出する。

Step 4: 獲得した手がかり表現から、それに係る節を取得する。

Step 5: Step 2 から Step 4 を新たな手がかり表現と共通

頻出表現が獲得されなくなる、もしくは、予め定めた回数まで繰り返す。

獲得した共通頻出表現・手がかり表現は形態素解析し、名詞と接頭詞、動詞の場合のみ結合した複合語とする。

3.4 Word2vec のモデル生成

債券市場における拡張文書データから金融市場特徴語候補を自動で取得するため、Word2vec のモデルを債券市場における拡張文書データを用いて作成する。Word2vec は単語の分散表現(ベクトル)を得る手法である。単語をベクトルで表現することで、単語間のコサイン類似度を計算することができ、意味のよく似た単語(類似度が高い単語)を抽出することができる。Word2vec のモデル生成時の学習データによって、同じ入力語であっても出力される類似度が高い単語が異なるため、債券市場における拡張文書データを学習データとすることで、債券市場における特徴的な語が出力されることを期待する。なお、学習データは形態素解析したのち、連続する品詞や接頭詞の場合は結合し、複合語として作成した。以下に、債券市場における金融市場特徴語候補の例を示す。

民間需要, 世界景気, インフレ懸念, 膨らむ,
ユーロ買い, 民間金融機関, 推し進める, 統計,
デフレ圧力, 引き下げられ, 跳ね上がる, 上昇基調,
景気対策, 積み上げる, 収益性, 価格変動リスク

4. 金融市場特徴表現の獲得

酒井らの手法[5]を用いて獲得した手がかり表現と抽出した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。表現獲得の概要を以下に示す。

Step 1: 獲得した共通頻出表現と手がかり表現の中から、手がかり表現のみを抽出する。

Step 2: 金融市場特徴語候補と抽出した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 3: 生成した文字列が表現として適しているか判定を行い、文字列を金融市場特徴表現として獲得する。

Step3 における生成した文字列が表現として適しているか

の判定は、生成した文字列が拡張文書データに含まれているかどうかで行い、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。以下に獲得した債券市場における金融市場特徴表現の例を示す。

債券価格が上昇し、値動きが続くと、歳出が多い、
国債金利が上がる、国内景気が回復、減少が加わる、
設備投資が増えると、国内需要が堅調、成長の続く、
債券売りが加速、債券相場が安定する、負債が減る、
利回り低下が加速した、利回り低下が目立った

5. 景気要因・債券要因への分類

「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、債券市場における拡張文書データを、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。

本章では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。なお、学習の精度を上げるため、拡張文書データのうち日本経済新聞の記事のみを利用した。

5.1 分類手法の概要

Step 1: 深層学習モデルにより、拡張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: 深層学習モデルにより、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。

5.2 使用した深層学習モデルと学習の手順

概要の Step 1 および Step 2 で使用した深層学習モデル

は Bidirectional LSTM と 多層パーセプトロン(MLP)を
組み合わせたモデルを用いた. 学習モデルを図 1 に示
す.

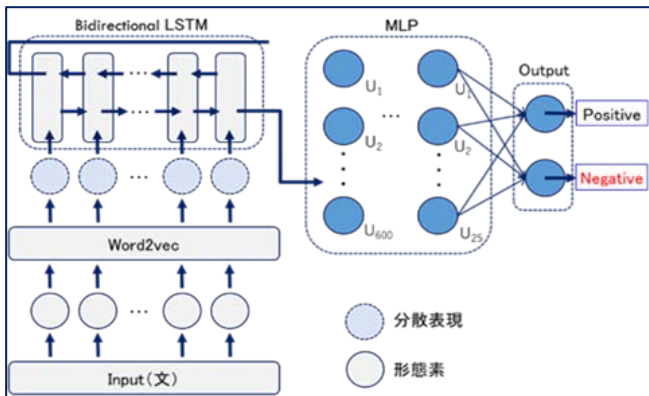


図 1 Bidirectional LSTM + 多層パーセプトロン(MLP)

学習は以下の手順で行う.

Step 1: 学習データの文を形態素解析し, 各語の分散表
現 (300 次元) を LSTM に入力する.

Step 2: 文頭から文末までのベクトルと文末から文頭まで
のベクトルを組み合わせ, 多層パーセプトロン(MLP)
の入力層 (600 ユニット) に入力する.

Step 3: MLP の出力層は2ユニットとなっており, それぞ
れ, positive, negative の尤度を出力する.

MLP の中間層は 9 層(ユニット数 300 が 3 層, ユニット数
100 が 3 層, ユニット数 25 が 3 層)である. エポック数は
Step 1 では 2 エポック, Step 2 では 5 エポックである. ま
た, 分散表現を得るために Word2vec を使用し, そのデー
タには日本経済新聞記事 10 年分を用いた.

5.3 学習データとテストデータの作成

概要の Step1, Step2 における深層学習のための学習
データは, 以下の手法にて自動生成した. 学習データを
自動生成することにより, 人手で作成することは難しい量
の学習データを用いて学習モデルを学習することができ
る. Step 1 における学習を行うために作成した学習データ
の詳細を以下に示す.

- 正例データ(下記の条件を全て満たす文)
 - 「景気」または「債券」を含む

- 手がかり表現を含む

- 負例データ(下記の条件を全て満たす文)
 - 「景気」とその分散表現の類似度が高い単語(「景気」
を含めて20個)を含まない
 - 「債券」とその分散表現の類似度が高い単語(「債券」
を含めて20個)を含まない
 - 手がかり表現を含まない
 - 共通頻出表現を含まない

作成した正例と負例のデータを合わせたものを学習デー
タ, 債券市場における拡張文書データをテストデータとし
た. そして, テストデータにおいて, ポジティブの尤度が
0.7 より大きい文を景気要因または債券要因に関する文
として抽出した.

概要の Step 2 における学習を行うために作成した学習
データの詳細を以下に示す.

- 正例データ
 - Step1 における学習データの正例において「景気」を
含む文
- 負例データ
 - Step1 における学習データの正例において「債券」を
含む文

作成した正例と負例のデータを合わせたものを学習デー
タ, Step 1 で抽出したデータをテストデータとした. テス
トデータにおいて, ポジティブの尤度が 0.7 以上の文を「景
気要因に関する文」, ネガティブの尤度が 0.7 以上の文を
「債券要因に関する文」として抽出した.

6. 金融市場特徴表現への極性付与

4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, 債券要因, それぞれを対象とした極
性の付与を行う.

6.1 金融市場特徴表現への極性付与の概要

Step 1: 獲得した手がかり表現(「が好調」「が不振」など)
を抽出し, その中から手がかり表現単体で考えたとき
に, 極性が明らかなものに対して人手で極性(ポジテ

イブ、ネガティブ)の付与を行う。

Step 2: 5章で抽出した景気要因に関する文の集合から、文中に4章で獲得した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 Step1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: 深層学習モデルを用いて、テストデータに極性(ポジティブ、ネガティブ)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と尤度をもとに極性の付与を行う。

Step 6: Step 1 から Step 5 までを債券要因においても同様に行う。

6.2 手がかり表現への極性付与

3章において、酒井らの手法[5]を用いて、金融市場特徴語候補を抽出するための手がかりとなる表現を債券市場における拡張文書データから自動的に獲得を行った。

抽出した手がかり表現の中から手がかり表現単体で考えた時に、ネガティブまたはポジティブが明らかなものに対して極性を人手で付与する。なお、景気要因と債券要因では付与する極性が異なるため、それぞれに対応する極性付与を行う。例えば手がかり表現「が懸念されている」は、景気要因ではポジティブ、債券要因ではネガティブになる。

6.3 学習データとテストデータの作成

5章で分類した景気要因に関する文または債券要因に関する文において、4章で獲得した金融市場特徴表現が文中に含まれている文のみを抽出し、これらをテストデータとする。さらに、テストデータの中から人手で極性付与した手がかり表現が含まれている文にはその極性を付与する。なお、複数の手がかり表現が含まれる場合には最後に出現する手がかり表現の極性を文に付与し、極性が付与された文をまとめて学習データとする。これを景気要因と債券要因のそれぞれで作成する。

テストデータへの極性付与は5章で示した深層学習モ

デルで行っており、エポック数は景気要因と債券要因ともに10エポックである。以下に極性付与された文の例を示す。

- ・ 景気要因に関する文でネガティブとなった例
第二に、ここまでの景気回復の「余熱」効果が期待される
- ・ 債券要因に関する文でポジティブとなった例
市場では債券買いが進み、円高・ドル安に動いている

6.4 金融市場特徴表現への極性付与

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。具体的には、ある金融市場特徴表現を含む文の尤度の平均をもとめ、それを金融市場特徴表現の極性の尤度とする。さらに、ネガティブ、ポジティブの尤度のどちらかが0.7以上の表現のみを抽出する。

金融市場特徴表現の極性付与において、景気要因での例を表1、債券要因での例を表2にそれぞれ示す。表1の極性は債券における景気要因の極性であり、景気が悪化すると債券市場における極性は positive が付与されることになる。

表1 景気要因における金融市場特徴表現への極性付与

極性	金融市場特徴表現	negative 尤度	positive 尤度
positive	景気への悪影響を	0.013	0.987
positive	景気回復が遅れる	0.219	0.782
negative	状況が上向く	0.999	0.001
negative	相場が底堅かった	0.998	0.001

表2 債券要因における金融市場特徴表現への極性付与

極性	金融市場特徴表現	negative 尤度	positive 尤度
positive	円買いが膨らんだ	0.159	0.842
positive	入札が堅調	0.091	0.909
negative	債券売りが進んだ	0.996	0.005
negative	国債保有が縮小	0.998	0.002

7. 評価

本手法についての評価は、景気要因における極性と債券要因における極性からそれぞれランダムに抽出した50個の金融市場特徴表現を対象として行った。なお、評価対象となる金融市場特徴表現の極性は、付与されている極性が正しいかどうかで判断した。評価結果を表3に示す。

表3 評価結果

要因	精度(%)
景気	80
債券	58
全体	69

8. 考察

金融市場特徴表現における極性の精度は、景気要因においては80%と高い精度を得ることができた。一方、債券要因においては58%と景気要因に比べて高い精度を得ることができなかった。これは拡張文書データから債券要因に関する文を抽出する際に、景気要因に関する文も一部混在してしまっているからだと考えられる。

景気要因と債券要因の両方で出現する金融市場特徴表現において、それぞれの要因で極性が正しく付与できていれば、整合性があるといえる。表4にどちらの要因にも出現した金融市場特徴表現の極性ごとの例を示す。

表4 景気要因と債券要因の両方で出現した極性ごとの金融市場特徴表現

	景気要因 ポジティブ	景気要因 ネガティブ
債券要因 ポジティブ	・円買いが広がる ・ドル安が続くと	・投資が堅調な ・市場が堅調
債券要因 ネガティブ	・不安が強まり ・期待が後退	・収益が回復し ・業績が堅調な

表4から、どちらの要因でもポジティブとなっているものは、債券市場にとって良い表現になっていることが分かる。どちらの要因でもネガティブとなっているものは、景気にとって良い表現となっている。景気要因でポジティブ、債券

要因でネガティブとなっているものは、状況が悪い、もしくはこれから悪くなることを示唆する表現となっている。景気要因でネガティブ、債券要因でポジティブとなっているものは、状況が良い、もしくは良くなることを示唆する表現となっている。債券市場においては、景気にとってネガティブな表現はポジティブな表現として考え、債券にとってポジティブな表現はそのままポジティブな表現として考えることができるので、いずれも整合性が取れている。

9. 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、日経 QUICK 月次調査と日本経済新聞の記事を用いて、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。その結果、極性の精度は景気要因で80%、債券要因で58%、全体で69%となり、比較的良好な結果を得ることができた。

今後の課題は、債券要因における極性の精度を向上させることである。

参考文献

- [1] 伊藤友助, 酒井浩之, 北島良三, 末廣徹, 稲垣真太郎, 木村柚里: 株式・債券・為替ごとの金融極性辞書の自動構築, 第24回金融情報学研究会, pp.211-218, 2020.
- [2] Tomoki Ito, Hiroki Sakaji, Kota Tsubouchi, Kiyoshi Izumi, Tatsuo Yamashita, Text-Visualizing Neural Network Model: Understanding Online Financial Textual Data, Advances in Knowledge Discovery and Data Mining. PAKDD 2018. Lecture Notes in Computer Science, Springer, vol 10939, pp 247-259, 2018.
- [3] 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.
- [4] 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.
- [5] 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.