2G5-OS-25b-5

金融市場ニュースの分散表現学習による辞書作成と金融市場分析

Constructing Financial Dictionaries through Distributed Representation Learning

片倉 賢治[*1]
Kenji Katakura

高橋 大志[*1]
Hiroshi Takahashi

[*1] 慶應義塾大学 大学院経営管理研究科
Graduate School of Business Administration, Keio University

In recent years, a method of machine learning technology has been advancing. Getting the distributed representation of words by the Neural Probabilistic Language Model (NPLM) has attracted interest. In this study, we attempt to create a new dictionary on the basis of ones widely used in finance. In this analysis, we use News Feed Direct of Thomson Reuters Corporation provides (NFD) that focus on the market trends of the Japanese stock market in order to make the dictionary. As a result of analysis, we could succeed in creating a new financial dictionary. However, we also found problems to be solved to get better performance.

1. はじめに

近年, 金融市場に対する関心が高まっている. 世界的な金融危機以降, 先進諸国を中心とした経済金融政策運営なども大きな転換点を迎えており, 個々の金融機関の健全性を確保するだけでは, 金融システム全体としての安定を必ずしも実現できるわけではないといった主張などもみられる. また, 同時に金融を分析する手法への関心も高まっている. とりわけ, 情報処理技術の向上等を背景として, 金融市場を取り巻く情報が年々飛躍的に増大しており, 従来からの分析手法に加え, 新たな分析手法に対する要望も高まっている.

株式や債券などといった資産価格を対象とした分析については, 従来より, 現実の市場データを用いた数多くの実証分析が報告されている. リスク資産を対象とした分析においては, 例えば, 合理的な投資家や摩擦のない市場等を仮定し導出される資本資産価格評価モデル(CAPM:Capital Asset Pricing Model)や, Fama-French の 3 ファクターモデルといった手法が広く用いられている[Sharpe 1964][Fama 1993]. また, 昨今, 金融市場を取り巻く, 企業の開示情報やニュース, マイクロブログといった広く利用可能で大規模な言語情報を分析対象とした新たな手法による研究も盛んに行われている. 例えば, ニュース記事を機械学習手法の一つ SVM(Support Vector Machines)によって分類し, 株価動向に関して分析を行った研究[Schumaker 2009]や, 深層学習(Deep Learning)と呼ばれる多階層ニューラルネットワークモデルの一つでもある RNN-RBM (Recurrent Neural Networks Restricted Boltzmann Machine)を用いて,時間的に変動する株価の上昇, 下落を予測した研究[吉原 2014]等, 国内外問わず数多くの研究報告がなされている.

言語情報を用いた分析において, 採用する辞書やその精度は, 重要な要素の一つである. SEC(米証券取引委員会)に提出された 10-Ks(年次報告書)を基にファイナンス分野の言語情報に特化した辞書を作成した研究では, 心理社会学で広く用いられている辞書 H4N(Harvard-IV-4 TagNeg)を用いて分析をしたものと比較して, 誤分類による影響が緩和され, 説明力が向上したとの報告が行われている[Loughran 2011]. また, これらファイナンス用辞書を金融市場の分析に活用した報告なども行われている[Yamashita 2013]. このように, 分析に採用する辞書とその精度は, 重要な要素の一つであり, より優れた辞書・単語群の構築の意義は大きい.

これらを背景とし, 本研究では, 金融市場ニュースからより優れたファイナンス用辞書の作成, そして, その辞書を用いた金融市場の分析を試みる. 具体的には, 金融市場ニュースの言語情報を CBOW(Continuous Bag-of-Words)という新たな手法で学習した結果とこれまでに提案されているファイナンス用辞書[Loughran 2011]をベースとして用い新たな辞書を作成し, その辞書を用いて金融市場との関係性の分析を試みる. なお, 本分析では, 日本の株式市場を対象とした英語ニュース記事群を分析対象とした.

本稿の構成は以下の通りである. 次節において, 本稿で採用する新たな辞書の作成手法である CBOW の概略について説明を行った後, 分析に用いるデータおよび分析方法について説明を行う. 次いで, 分析結果, 考察を示した後,本稿のまとめを示す.

2. CBOW (Continuous-Bag-of-Words)

CBOW は, 近年注目を集めている新しい機械学習手法であり,単語の分散表現を高精度で獲得できるという特徴を有している. 分散表現とは, 単語を K 次元で一意に表現するという 1-of-K 符号化によって得られるベクトルをより低次元で表現したものであり, 意味が近しい単語同士はそのベクトル距離が近くなるような性質を有する表現を指す.

従来から言語情報の分析には, 文章中に現れる単語を扱う BOW(Bag-of-words)によって表現する手法が一般的であったが, 順序性の欠如, 扱う単語数によって膨大な次元数となる等の欠点があった. しかしながら, CBOW によって学習した分散表現

連絡先:片倉 賢治, 慶應義塾大学 大学院経営管理研究科,
〒223-8526 神奈川県横浜市港北区日吉4−1−1

はこれらの課題を克服し，更に精度も向上するとの報告が行なわれている[Mikolov 2013a] [Mikolov 2013b] [西尾 2014].

次図 1 は，CBOW の分散表現の獲得方法の概略を示したものである．図の左から，Input, Projection, Output となっている．CBOW においては，注目する単語 w(t)の前後の単語群 w(t-2), w(t-1), w(t+1), w(t+2)から構成される BOW を入力とし，注目する単語 w(t)を出力するニューラルネットワークの学習により，分散表現が獲得される[1]．また,分散表現（単語ベクトル）を学習した結果を基に，各単語と距離が近い単語群を出力することが可能である[2]．なお，本分析では，コサイン距離を採用した．

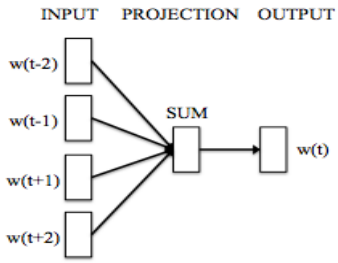

図 1: CBOW の概略.

3. データ

本稿では，金融市場ニュースおよびベースとする辞書，更に，市場関連データを用いる．

金融市場ニュースに関しては，世界で最も広く知られたニュース提供会社の一つである Thomson Reuters 社により提供されているニュースを採用した．具体的には，世界のマーケット動向に関するニュースとして，Thomson Reuters 社提供の News Feed Direct (NFD)を用いた．NFD は，News Scope Direct としても知られており，ニュースのヘッドラインや経済イベントを極小の遅延で配信し，発表時刻もミリ秒単位で保持している等，分析に適した特徴を有している．

ベースとする辞書は，前述のファイナンス辞書[Loughran 2011]を採用し，Positive な単語，Negative な単語の一覧をそれぞれ Web ページから入手した．

市場関連データは，日経 NEEDS および Thomson Reuters Datastream より入手した．本分析ではファクターリターンを対象とした分析を行うため，ファクターモデルに関するデータ（以下，FF ファクター）は，久保田(2007) に従い、日本における東証 1 部，東証 2 部から構成される銘柄から算出した 3 ファクター（マーケットとの関係性を表す Rm - Rf，企業規模に係る小型株効果を表す Small minus Big：SMB，割安株効果を表す High minus Low：HML）データを用いた．

サンプル期間は，2003 年 1 月 1 日～2012 年 7 月 31 日とした．NFD は，世界各国の市場を対象としたニュースが含まれており，ニュースの言語も英語，フランス語，ドイツ語，日本語などをはじめ多岐にわたる．このような膨大な量のニュースの中から本分析では，とりわけ，日本市場及び日本企業に関する英語ニュース記事 411,531 件を対象として分析を行った．ニュース記事は，1,349 万行，9,265 万単語を含み，対象記事の内，日本企業に関連する記事は，363,970 件，該当する企業数は 308 件であった．

4. 分析方法

本研究では，金融市場ニュースから分散表現を学習し，新たなファイナンス辞書の作成，分析を試みる．

そこで，分析に先立ち図 2 に示すシステムを構築した．システムは主に 4 つのモジュール，(1) NFD データベースから記事を抽出，(2) 抽出した記事群の整形，(3) ファイナンス辞書から単語を読み込み，distance モジュール（インプットした単語とコサイン距離が近い単語を出力）をバッチ形式で連続実行，(4) 出力された単語リストを読み込み，日次 NFD での出現頻度を日次カウントし結果を CSV ファイルで保存，をそれぞれ用意した．

図 2 中の左部分は，金融ニュース (NFD)，既存のファイナンス用辞書(Financial Dictionary)を示し，本分析では，これら情報を基に新たな辞書を作成した後，市場データとの関係性を分析する．

具体的な分析方法は，次のとおりである．

(1) 全期間のニュースデータ(NFD)を 1 つのファイルに抽出する．
(2) 分析対象のニュースデータ(NFD)に対して,必要な前処理を行う．(例：URL 及び E メールアドレスの除外処理を行う．全ての大文字を小文字に変換する等.)
(3) ニュースデータ(NFD)ファイルを分散表現学習モジュールに入力し，分散表現を学習する．本分析では，学習手法は階層化ソフトマックスとし，考慮する文脈サイズは 5（w(t-5)～w(t+5)）とした．
(4) 分散表現学習モジュールの出力結果を，距離算出のための入力として実行する．
(5) 既存辞書の Negative 単語を距離算出モジュールに順次入力し，分散表現のコサイン距離が近い単語群の一覧を取得する．
(6) ファイナンス辞書で定義済み単語 及び 獲得した新単語に対して日次 NFD 出現頻度を計上し，センチメントスコアとして算出する．
(7) センチメントスコアとマーケットデータの関係性を分析する．

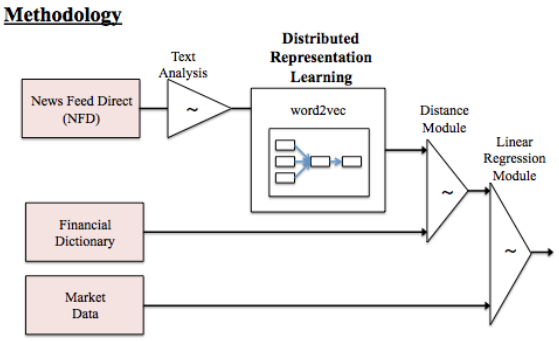

図 2: システム概要図.

5. 分析結果

本稿では，はじめに既存辞書の Positive 単語一覧をベースとした分析を行った後，Negative 単語一覧をベースとした分析を

[1] 本稿においては，分散表現学習において，word2vec と呼ばれるモジュールを採用した．https://code.google.com/p/word2vec/

[2] 本稿においては，距離算出において，distance モジュールを採用した．

行い，新たな辞書の作成，マーケットとの関係性の分析を行った．

5.1 Positive 単語

分析の結果，既存辞書の単語全 354 件中，金融市場ニュースに出現した単語は 335 件，出現しない単語は 19 件であることがわかった．次いで，距離算出モジュール(distance モジュール)に，この出現した単語のみを入力し，コサイン距離の近い（本稿では閾値を 0.7 とした）単語リストを獲得した．結果，コサイン距離が閾値以上の単語数は 35 単語であり，それらのうち，既存辞書に含まれていない新たな単語は 20 単語であった．

5.2 Negative 単語

同様に，既存辞書の単語全 2,329 件中，金融市場ニュースに出現した単語は 1,933 件，出現しない単語は 396 件であることがわかった．次いで，コサイン距離の近い単語リストを同様に獲得した．結果，コサイン距離が閾値以上の単語数は，67 単語であり，それらのうち，既存辞書に含まれていない新たな単語は 45 単語であった．

5.3 新たに抽出した単語群

それぞれの新たに抽出された単語群は，重複を除くと 54 件あった．この中には，新たに drop，fell など市場全体に対して Negative な極性を有する可能性のある単語と，rose，climbed などといった Positive な極性を有する可能性のある単語を含むものであった．

5.4 センチメントスコアの算出

既存の辞書および新たに作成した辞書を用いて，日次 NFD に出現する単語頻度を算出しセンチメントスコアとした．

図 3 および図 4 は，それぞれ算出したセンチメントスコアの時系列データ（Positive および Negative スコア）の自己相関を示したものである．図 3，図 4 の横軸は東証営業日を基準とした時間的な差異(Lag)，縦軸は自己相関係数を表す．また，図の青破線は 95%信頼区間を示している．図 4 より，Negative スコアは 1～20 営業日(1 ヶ月間)に渡って自己相関係数の値が大きくなっており，更に 50～70 営業日（3 ヶ月程）に再び自己相関が大きくなっていることを確認できる．Positive スコアについては Negative スコアと比較すると自己相関は小さいものの同様の傾向を確認できる(図 3)．本結果は，ニュースより見積もられた各スコアが，時系列相関を有する可能性を指摘するものであり，興味深い結果を示すものである．

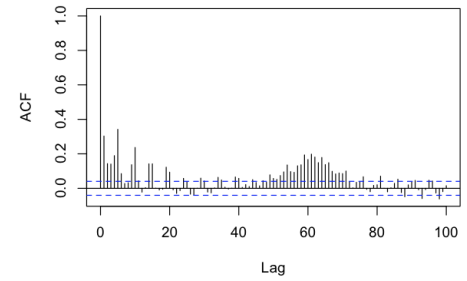

図 3: Positive スコアの自己相関．

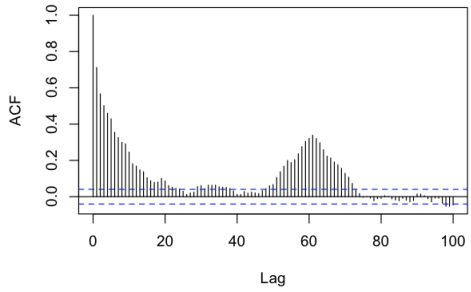

図 4: Negative スコアの自己相関．

5.5 マーケットとの関係性

表 1 は，マーケットファクターとセンチメントスコアの関連性を分析した結果の抜粋を示したものである．本分析では，同時点の関係性について分析を行った．表中においては，従来の辞書と新しい辞書（従来単語に新単語を加えた単語群），それぞれの辞書を用いた結果を示しているが，従来のファイナンス辞書に比べ新辞書によって算出したスコアは，より強い関連性を有していることを確認できる．更に，Negative スコアはマーケットファクター（Rm - Rf）に，Positive スコアは企業規模ファクター（SMB）に，それぞれ強い関係性を有していることを確認できる．

表 2 は，企業規模および時価総額を基に企業を分類し構築したポートフォリオとセンチメントスコアとの関連性を分析したものである．なお，表中の BL，BM，BH は，東証 1 部および東証 2 部から構成される銘柄の中で，時価総額が中央値以上かつ時価総額に対する自己資本比率が 30%以下，30~70%，70%以上の企業のポートフォリオ示したものであり，SL,SM,SH は,時価総額が中央値より小さくかつ時価総額に対する自己資本比率が 30%以下，30~70%，70%以上の企業のポートフォリオを示したものである.表より，Positive スコアは特に時価総額の大きい企業群のファクターに対して正の相関を有していることを確認できる．

表 1: センチメントスコアとファクターリターンとの関連性．

'***' P< .001 , '*' P< .05 , '.' P< .1

被説明変数	説明変数	回帰係数	T値	調整済み決定係数
Rm – Rf	Negative(従来)	-.00047 ***	-4.109	.00672
Rm – Rf	Negative(新辞書)	-.00501 ***	-7.482	.02286
SMB	Positive(従来)	-.00059 **	-3.216	.00395
SMB	Positive(新辞書)	-.00047 ***	-3.307	.00421

表 2: センチメントスコアと各ポートフォリオとの相関係数．

相関係数	Positive
SL	.00031
SM	.01317
SH	.01121
BL	.03038
BM	.04368
BH	.03821

6. 考察

本研究では，既存のファイナンス用辞書をベースとして CBOW によって新たな単語群の抽出を行った．抽出する際，CBOW のパラメータを固定して分析を行ったが（例えば，文脈サイズを 5 に固定），より改善した手法による詳細な分析は今後の課題として挙げられる．また，辞書に定義されている単語群は

マーケット全体との関係性に注目しポジティブ・ネガティブに分類された単語群であることから，より細分化した業界・個別企業との関係性に注目した単語群の定義についても今後の課題である．

7. まとめ

本稿では，金融市場ニュースの分散表現学習による新たな辞書作成を試みた．分析の結果，既存辞書には含まれない新たな単語群を抽出することができた．更に，新たな単語群を基に分析したところ，従来辞書と比較してファクターリターンとの強い関連性があることを見出した．更に，本分析では，ニュース記事より算出したスコアが，時系列相関を有する可能性があることなど，興味深い結果を見出すことができた．

一方，本稿の分析は，今後更なる精度の改善の余地があることから，より詳細な分析は今後の課題である．具体的には，抽出する単語群の閾値であるコサイン距離の範囲の調整，CBOWの学習パラメータ調整，他のニュース媒体における新たな単語群の抽出や比較，他国のマーケットとの関係性の比較などが，今後の課題として挙げられる．

参考文献

[Sharpe 1964] Sharpe, W. F.: Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *The Journal of Finance*, 19(3), 425-442, 1964.

[Fama 1993] Fama, E. F., & French, K. R.: Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56. , 1993.

[Schumaker 2009] Schumaker, R. P., & Chen, H.: Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 12. , 2009.

[吉原 2014] 吉原輝, 藤川和樹, 関和広, & 上原邦昭.: 深層学習による経済指標動向推定. *人工知能学会全国大会論文集*, *28*, 1-4. , 2014.

[Loughran 2011] Loughran, T., & McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10 -Ks. *The Journal of Finance*, *66*(1), 35-65. , 2011.

[Yamashita 2013] Yamashita, Y., Joutaki, H., & Takahashi, H.: Analyzing the influence of headline news on the stock market in Japan. *International Journal of Intelligent Systems Technologies and Applications*, *12*(3), 328-342. , 2013.

[Mikolov 2013a] Mikolov, T., Chen, K., Corrado, G., & Dean, J.: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. , 2013.

[Mikolov 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.: Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111-3119). , 2013.

[西尾 2014] 西尾泰和.: word2vecによる自然言語処理, オライリージャパン, 2014.

[久保田 2007] 久保田敬一, & 竹原均.: Fama-French ファクターモデルの有効性の再検証. *現代ファイナンス*, (22), 3-23. , 2007.