1J4-OS-13a-5

再帰的ニューラルネットワークによる感情分析モデルを用いた株価動向予測

Applying a Sentiment Analysis Model of Recursive Neural Network for Stock Price Prediction

秋田 諒 *1 Ryo Akita　吉原 輝 *1 Akira Yoshihara　関 和広 *2 Kazuhiro Seki　上原 邦昭 *1 Kuniaki Uehara

*1神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

*2甲南大学知能情報学部知能情報学科
Department of Information Science and Systems Engineering, Konan University

This study proposes an approach to predicting stock price movements caused by news events. Previous works on news-based stock price prediction often use bag-of-words as features which cannot represent the word order of sentences. In our approach, we employ the Recursive Neural Network that models semantic compositionality so as to accurately analyze implications of the news texts for stock prices. Using Reuter's news archives, we evaluate the validity and effectiveness of the proposed approach for stock price prediction.

1. はじめに

株価は株式市場が開いている間，常に変動している．変動の要因は様々であり，日経平均など市場全体に影響する要因や，企業の合併や業績修正などの個別企業に影響する要因などがあげられ，これらはニュースとして報じられる．しかし，報じられた全てのニュースに投資家が目を通し，その中から自身に関係のある企業のニュースを取り上げ，投資に有益な情報かどうかの判断を下すのは困難である．そこで，Hagenau ら [Hagenau 12] や Xie ら [Xie 13] など，近年では自然言語処理技術を用いてテキスト情報を自動的に解析し，株価を予測する研究が行われている．

これらの先行研究は，bag-of-words によりテキストを表現している場合が多い．しかし，bag-of-words による表現では，語順が考慮されないという欠点がある．例えば「神戸鋼：連結、14 年 3 月当期黒字転換、15 年 3 月期予想 28.8%減」というニュースが報道された場合，bag-of-words では,「神戸鋼」,「連結」,「14 年」,「3 月」,「当期」,「黒字」,「転換」,「15 年」,「3 月期」,「予想」,「28.8%減」というキーワードの語順が無視され,「14 年」と「15 年」のどちらが「黒字」であるかなどの情報が失われてしまう．よって，投資に有益な情報を抽出する事ができない．

そのため本研究では，ニュース記事を単語の係り受け構造が考慮された構文木で表現する．そして，テキストの係り受け構造を捉えるため，再帰的ニューラルネットワーク（Recursive Neural Network; RNN）を感情分析へと応用したモデル [Socher 13] を用いる．この感情分析モデルを援用することによって，先ほどの例で挙げたニュースに対しても,「14 年 3 月当期黒字転換」という単語列を考慮することができ，bag-of-words の問題点を解決することができる．このようにテキスト情報の語順や係り受け構造を捉えて表現することで，株価動向の予測を行う．

連絡先: 秋田 諒，神戸大学大学院システム情報学研究科，akita@ai.cs.kobe-u.ac.jp

2. 関連研究

2.1 Word Embedding

Word Embedding とは，単語が持つ構文的・意味的な情報をベクトルを用いて表現する手法で，ニューラルネットワーク言語モデル [Bengio 03] 等によって学習することで得られる．

単語を表現するベクトルの次元数を d，語彙数を $|V|$ とするとき，$d \times |V|$ の行列を辞書行列と呼ぶ．また，ある単語ベクトルは，その単語のインデックスのみが 1 で残りの要素が全て 0 である $|V|$ 次元のベクトルと辞書行列の行列演算によって辞書行列から取り出すことが可能である．

こういった単語のベクトル表現は Distributed Word Representation などとも呼ばれ，意味解析や文書分類など様々な分野での応用が期待されている．例えば，後述する再帰的ニューラルネットワーク（RNN）によって，Word Embedding から，文が持つ構文的・意味的な情報をベクトルで表現するモデルが提案されている．

2.2 テキスト情報を用いた株価動向推定

Hagenau ら [Hagenau 12] は，DGAP，EuroAdhoc と呼ばれるドイツとイギリスの企業報告書データを基に，サポートベクターマシン（SVM）により当日の株価の始値と終値の差分の正負の予測を行った．連続する 2 単語（バイグラム）や近傍の 2 単語の組み合わせに対してカイ二乗検定量によって素性選択を行い，過学習を削減することで分類精度を向上させた．

和泉ら [和泉 11] は，共起解析（co-occurrence analysis），主成分分析（principal component analysis），そして回帰分析（regression analysis）を組み合わせてテキストを解析することで経済動向の予測を行う CPR 法を提案した．一方，和泉らの研究では入力テキストが日本銀行の金融経済月報であったのに対し，藏本ら [藏本 13] は CPR 法を拡張し，金融経済月報に比べて文章の形式が定まっていない新聞記事を用いて，市場動向推定を行った．藏本らの研究では，TOPIX や日経平均といった市場平均株価の騰落を予測し，63.7% という精度が報告されている．

3. RNN を用いた株価動向予測

本章では，再帰的ニューラルネットワーク（RNN）を感情分析に応用したモデルについて説明し，そのモデルを株価動向予測に援用する方法について述べる．

3.1 RNN の概要

Word Embedding による単語のベクトルを用いて単語列や文を表現することで，文の意味のベクトル表現を試みる研究が盛んに行われてきた．例えば，Mitchell ら [Mitchell 10] は，2 語からなる単語列の意味的な類似度を，単語列を構成する単語のベクトルの加算や乗算によって計算した．しかし，固定長の単語列をベクトルで表現することが可能であっても，実際の文に含まれる単語数は，文によって異なる．つまり，Word Embedding によって文を表現するためには，可変長の入力に対応したモデルが必要になる．そこで，Socher ら [Socher 13] は，深層学習のモデルの一つである RNN を用いて，可変長の入力に対応した感情分析モデルを提案した．

Socher らの手法では，文を構文解析器にかけ，2 分木の構文木を作成する．その構文木に基づき，全てのノードに d 次元のベクトルと一つのラベルを付与する．ラベルとは分類問題におけるラベルで，感情分類であれば,「Positive」や「Neutral」などの感情ラベルである．感情分類における例を図 1 に示す．葉ノードである $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ のベクトルが表現するのは，Word

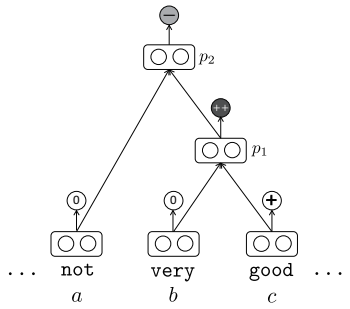

図 1: RNN モデルの概念図.

Embedding と同様に単語の意味的な情報であり，それ以外のノード $\boldsymbol{p}_1, \boldsymbol{p}_2$ が示すのは，自身の 2 つの子ノードが示す単語（列）を連結させた単語列が持つ意味的な情報である．また，ノードに付与された「0」（Neutral）や「+」（Positive）などは，単語（列）の感情ラベルを示す．親ノードのベクトル $\boldsymbol{p}_1$ は，子ノードのベクトル $\boldsymbol{b}, \boldsymbol{c}$ を連結させた $2d$ 次元のベクトルを入力層とし，$\boldsymbol{p}_1$ を隠れ層としたニューラルネットワークのように以下の式で計算される．

$$\boldsymbol{p}_1 = f\left(\boldsymbol{W}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{c}\end{bmatrix}\right) \tag{1}$$

ここで f は活性化関数で，非線形関数 tanh を用いる．$\boldsymbol{W} \in \mathbb{R}^{d\times 2d}$ は重みで，入力層の $2d$ 次元のベクトルを d 次元に圧縮する役割を持つ．つまり RNN では，子ノードからボトムアップ式に計算することで，単語列が持つ意味的な情報を d 次元ベクトルで表現することが可能である．さらに，単語列ベクトルは構文木に基づいて構成されているため，構文的な情報までも表現している．また，構文木の根のベクトルは文に含まれる全ての単語を連結させた単語列，すなわち文そのものを表すベクトルである．これにより，いかなる文でも意味的・構文的な情報を持った固定長の次元 d のベクトルで表現することが可能である．

そのように得られたベクトルを隠れ層とみなし，ソフトマックス関数によるラベルの予測を次式により行う．

$$\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{W_s}\boldsymbol{p}_2) \tag{2}$$

$\boldsymbol{W_s} \in \mathbb{R}^{N\times d}$ は，ソフトマックス関数に対する重みである．N は，分類問題におけるラベルの種類を示す．この計算は，文だけでなく，全ての単語，単語列のベクトルに対して行われ，例えば感情分類のタスクの場合，その単語や単語列が持つ感情ラベルが予測される．

予測されたラベルと正解ラベルの誤差を最小化するためにパラメータを学習する．学習するべきパラメータは $\theta = (\boldsymbol{W}, \boldsymbol{W_s}, \boldsymbol{L})$ である．ここでの $\boldsymbol{L} \in \mathbb{R}^{d\times|V|}$ は，2.1 節で述べた辞書行列を指す．$\boldsymbol{y}^i \in \mathbb{R}^{N\times 1}$ をノード i の $\boldsymbol{y}$，$\boldsymbol{t}^i \in \mathbb{R}^{N\times 1}$ をノード i の正解ラベルのみが 1 で他の要素が 0 であるベクトルとし，それらのベクトルの要素のうち，j 番目の要素をそれぞれ y^i_j，t^i_j とする．そして，λ を正規化項の損失を調整するパラメータとするとき，θ は以下の式で表される交差エントロピー関数を最小化するように誤差逆伝播法で学習される．

$$E(\theta) = \sum_i \sum_j t^i_j \log y^i_j + \lambda\|\theta\|^2 \tag{3}$$

3.2 日本語に対する 2 分木の獲得

前節で述べたように，RNN は構文解析器によって得られた 2 分木に基づいた計算を行う．日本語の構文解析とは，主に分節間の係り受け構造を発見することであるため，本研究では日本語における係り受け解析器である CaboCha[*1] を用いた．CaboCha により得られた係り受け構造を 2 分木に変換したニュース見出しの例を図 2 に示す．

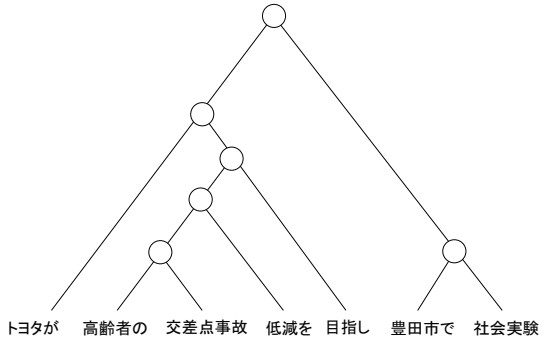

図 2: 2 分木で表現した日本語のニュース見出しの例.

3.3 本手法で用いる素性

3.3.1 予測対象

本研究では，記事の発行時刻の株価を基準値，企業の株価に影響を与える時間を 15 分以内と仮定し，その期間の株価の動向を予測した．予測対象は実際のデイトレーダーが株の売買をするときに意識している利食いや損切りと呼ばれるものである．利食いとは，自身の持ち株の株価が上昇したことを受けて，その株を売却し利益を得ることで，損切りとは，持ち株の株価が下落した場合，それ以上の損失を出さないために，早めに持ち株を売却することで損失を抑えることである．どちらも

*1 https://code.google.com/p/CaboCha/

株を購入する際にある閾値を設けて，その値を超えた時に売却することが多い．

例えば，100 万円分の株を購入する場合，利食いの閾値として 2%上昇の 20,000 円の利益が出た時に売却，損切りの閾値として 0.5%下落の 5,000 円の損失が出た時に売却のように決定する．今回利食い，損切りの閾値を ±1% として，15 分以内に 1 度でも株価が基準値から 1% 以上上昇すれば利食い（Up），1% 以上下落すれば損切り（Down），15 分間 ±1% 以上の変動がなければ安定（Neutral）として，3 値分類で予測を行う．

3.3.2 構文木のノードに用いるラベル

先に述べたように，RNN の感情分析モデルは，係り受け構造などを考慮したベクトルで文を表現できるという特徴を持つ．本項では，このモデルを株価動向推定に援用する方法について述べる．

本研究では，ニュース記事の見出しのみをテキスト情報として扱う．そして，企業のニュースが報じられた時，そのニュースの見出しが株価動向にどのような影響を及ぼすかについて，3.3.1 節で述べた 3 値に分類を行う．すなわち，実際にその見出し（構文木の根）が発行されたのちに起こる対象企業の株価の変動に当てはまるラベルの予測を目的とする．学習データの見出しに対しては，実際の株価動向に基づきラベル付けを行うことができる．しかし，3.1 節で述べたように，Socher らの手法では構文木のラベル付けを行う場合，文のみならず，単語や単語列に対してもタスクに応じた何らかのラベルを付与する必要がある．そこで本研究では，見出しに含まれる単語（列）に対して次のようにラベル付けを行う．ある企業の見出しに出現する単語（列）へのラベル付けを行う場合，対象企業の全ての見出しから対象単語（列）が含まれる見出しを抽出する．そして，それらの見出しに付与されたラベルを基に，ラベル毎の見出し数を集計し，最も多い見出し数となったラベルを対象単語（列）のラベルとする．また，最も多い見出し数となったラベルが複数あった場合，ラベルは Neutral とする．

例えば，神戸製鋼（銘柄コード : 5406）の「上方修正」という単語に対してラベル付けを行う場合，神戸製鋼について報じた見出しの中から「上方修正」という単語が含まれる見出しを抽出する．そして，それらの見出しのラベルを基に，ラベル毎の見出し数を集計する．その結果が「Up」が 7 件,「Neutral」が 1 件,「Down」が 0 件だった場合，この中で最も見出し数が多いラベルは「Up」なので，神戸製鋼の「上方修正」という単語には,「Up」のラベルを付与する．

このように，見出しだけでなく単語（列）に対しても，ラベル付けを行うことによって，RNN の感情分析モデルを株価動向予測へ利用することが可能になる．

4. 評価実験

4.1 実験データ

本研究では，ロイターの Machine Readable News をテキスト情報として用いた．期間は，2013 年 1 月から 2014 年 6 月までの 1 年半で，日本語で記載された記事の見出しのみを扱った．ロイターの見出しには，そのニュースに関連する企業の証券コードがタグ付けされている．本実験では，直接企業の株価に与える影響が強い見出しとして，東証一部に上場している証券コードのタグが付いており，かつ記事の発行時刻が東証一部の市場の取引時間内（9 時〜15 時）となっている記事の見出しのみを使用した．これらのうち，2013 年 9 月までの 5,054 件を訓練データ，2013 年 10 月の 330 件を検証データ，2013 年 11 月から 2014 年 6 月までの 4,763 件をテストデータとした．

3.2 節で述べた CaboCha では，係り受け構造を発見すると同時に，形態素解析も行っているため，本実験では，助詞や記号をストップワードとして取り除いた見出しを基に 2 分木を作成した．各データセットにおける，Up，Neutral，Down の見出しの数を表 1 に示す．

株価データとして，東証のウェブサイトからダウンロード可能 [*2] な歩み値データの中で東証一部に上場している企業のみに注目し，各企業の 1 分ごとの株価を求めたデータを用いた．歩み値とは，特定の株式銘柄について，いつどれだけの株数がいくらの株価で取引が成立（約定）したのかを示す値である．例えば，ある企業の株が，午前 9 時 2 分 10 秒に 402 円，午前 9 時 4 分 32 秒に 414 円で約定されたという歩み値データが存在した場合，この企業の株は，午前 9 時 3 分には取引が行われなかったことになる．よって，午前 9 時 2 分〜午前 9 時 3 分の株価は 402 円，午前 9 時 4 分の株価は 414 円のように毎分の株価を求める事ができる．

表 1: データセットにおける各ラベルの割合.

Dataset	train	dev	test	sum
Up	1,311	48	1,362	2,721
Neutral	2,617	240	2,340	5,197
Down	1,126	42	1,061	2,229
sum	5,054	330	4,763	10,147

4.2 実験結果

4.2.1 全企業に対する予測

提案手法を用いて，テストデータに含まれる全ての見出しに対して株価動向予測を行った結果を表 2 に示す．本研究では，誤差関数の最適化に AdaGrad [Duchi 11] を用いた．また，構文木のすべてのノードに割り当てるベクトルの次元数を 45，AdaGrad で用いる学習率の初期値を 0.1 とした．テストデータの見出しについて Up，Neutral，Down の数を集計し，最も数が多いラベル（今回は Neutral）に全ての見出しを分類したときの結果をベースラインとする．また，ベースラインの他に比較手法として，Naive Bayes と以下に述べる VecAvg を用いた．Naive Bayes は訓練データにおいて，出現した見出しが 40 個以下の単語を無視した bag-of-words を特徴量とした．VecAvg とは，RNN によって学習された Word Embedding を用いて，見出しを構成する単語の Word Embedding の平均を特徴量とし，ソフトマックス関数で予測を行った手法である．すなわち RNN と VecAvg の違いは，文の語順や構文構造などを考慮しているかという点である．また今回，Neutral に分類される見出しが多かったため，評価指標として精度（Acc.）だけでなく，マクロ平均により算出した F 値（F-measure）も用いた．

RNN は，全ての比較手法に比べて最も良い精度と F 値を実現した．

4.2.2 特定企業に対する予測

本手法の有用性を検証するために，データセット中で，報じられた見出しが多い上位 10 社に対して，先ほど学習したモデルで予測を行った．結果を表 3 に示す．なお，スペースの都合上，精度についての結果は省略しているものの，大まかな傾向は表 2 と変わらない．

*2 http://ec.tse.or.jp/

表 2: 株価動向予測結果.

	Acc.	F-measure
Baseline	49.1%	0.23
Naive Bayes	35.3%	0.33
VecAvg	50.2%	0.47
RNN	**53.8%**	**0.49**

表 3: 記事数上位 10 社に対する予測結果（F 値）.

	RNN	Naive Bayes	VecAvg	Baseline
Toyota	**1.00**	0.65	0.95	**1.00**
SoftBank	0.64	0.72	**0.84**	0.64
Sony	0.28	0.29	**0.54**	0.28
Sharp	**1.00**	0.62	0.44	0.63
Panasonic	**0.66**	0.21	0.34	0.32
Toshiba	**0.66**	0.21	0.54	0.31
Mitsubishi	**0.68**	0.49	0.36	0.24
Docomo	**1.00**	0.95	**1.00**	**1.00**
Kobe Steel	**0.49**	0.05	0.42	0.15
Tokyo Electric	**0.65**	0.54	0.49	**0.65**
All	**0.60**	0.33	0.49	0.30

t 検定による両側検定を行ったところ，RNN と VecAvg は Naive Bayes と比較して，有意水準 1% で有意差を確認することができた．このことから，RNN や VecAvg で用いた特徴量である Word Embedding は，Naive Bayes で用いた特徴量の bag-of-words に比べ，株価動向推定に有用であることが分かった．また，RNN はベースラインに比べて，有意水準 5% で有意差が確認できた．一方，VecAvg に対しての RNN の優位性は確認できなかった（$p = 0.17$）.

この原因を明らかにするため，テストデータにおける CaboCha の係り受け構造の出力結果を分析した所，図 3 のような誤った解析を行っている例が存在した．図 3 では,「買い」と「気配」という単語は複合名詞「買い気配」を構成する．しかしながら，CaboCha の解析では，これらの単語の係り受け関係を誤って認識し,「買い気配」という単語列が考慮されていない．すなわち，CaboCha の解析結果は「気配・パソコン事業の・売却交渉を・材料視」を「買う」のような解釈となっていた．そこで，このような見出しについて人手で構文木を作成し，再び RNN で予測を行ったところ，F 値の向上が見られた．この RNN の結果を基に再び VecAvg との差を検定したところ，有意水準 5%で有意差が確認された．

このことから，入力文の構文木が正しく認識できれば，RNN のような見出しの係り受け構造を考慮したモデルは，VecAvg のような係り受け構造を考慮していないモデルに比べて，株価動向推定に有用であると考えられる．

5. 結論

本研究では，テキスト情報であるニュース速報サービスの見出しが企業の株価に影響を与えると仮定し，自然言語文における語順や係り受け構造を表現できる RNN の感情分析モデルを援用して株価動向の予測を行った．

その結果，語順や係り受け構造を考慮していない他のモデルに比べ，RNN が最も良い精度および F 値を実現した．また，テストデータの中で記事数上位 10 銘柄に注目したところ，従来の研究で一般的に用いられている bag-of-words を特徴量とした Naive Bayes に対して，有意な精度向上が確認された．この結果から，bag-of-words に比べて，単語をベクトルで意味的に表現した Word Embedding が株価動向推定に対して有用である事が分かった．また，特徴量として Word Embedding が用いられているものの，構文構造は考慮されていないモデルに対しても，人手で修正した構文木を入力とする RNN で，F 値で有意な精度向上が見られた．今後の課題は，時間的な特徴量を考慮したモデルの拡張が挙げられる．

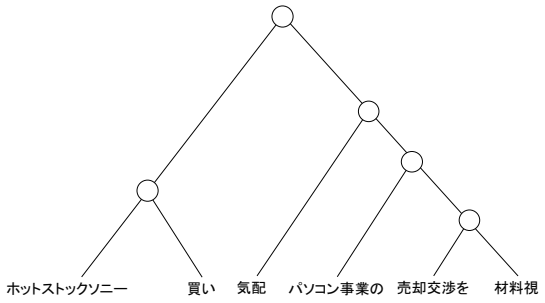

図 3: CaboCha による誤った解析結果の例.

参考文献

[Bengio 03] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C.: A neural probabilistic language model, *JMLR*, Vol. 3, pp. 1137–1155 (2003)

[Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization, *JMLR*, Vol. 12, pp. 2121–2159 (2011)

[Hagenau 12] Hagenau, M., Liebmann, M., Hedwig, M., and Neumann, D.: Automated news reading: Stock price prediction based on financial news using context-specific features, in *Proc. of the 45th Hawaii International Conference on Systems Science*, pp. 1040–1049 (2012)

[Mitchell 10] Mitchell, J., and Lapata, M.: Composition in distributional models of semantics, *The Journal of Cognitive science*, Vol. 34, No. 8, pp. 1388–1429 (2010)

[Socher 13] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank, in *Proc. of 2013 EMNLP*, pp. 1631–1642 (2013)

[Xie 13] Xie, B., Passonneau, R. J., Wu, L., and Creamer, G.: Semantic frames to predict stock price movement, in *Proc. of the 51st ACL*, pp. 873–883 (2013)

[和泉 11] 和泉 潔，後藤 卓，松井 藤五郎：経済テキスト情報を用いた長期的な市場動向推定, 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315 (2011)

[藏本 13] 藏本 貴久, 和泉 潔, 吉村 忍, 石田 智也, 中嶋 啓浩, 松井 藤五郎, 吉田 稔, 中川 裕志：新聞記事のテキストマイニングによる長期市場動向の分析, 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296 (2013)