自己注意機構付きLSTMを用いた景況感情報に基づく 金融文書の重要文抽出

Extraction of Important Sentences in Financial Documents Based on Business Confidence Information Using LSTM with Self-Attention Mechanism

山岡 周平 *1	小澤 誠一 *1*2	廣瀬 勇秀 * ³	飯塚 正昭 * ³
Shuhei Yamaoka	Seiichi Ozawa	Takehide Hirose	Masaaki Iizuka

*¹神戸大学大学院工学研究科 電気電子工学専攻 Department of Electrical and Electronic Engineering, 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 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 Sumitomo Mitsui DS Asset Management Company, Limited

Investment trust and fund management companies have accumulated a large number of visit records that were summarized by their analysts after conducting hearings against companies. Such visit reports include crucial information of companies such as companies' financial conditions and future strategies, which are used to estimate investment values of individual companies. However, it is not easy even for skilled fund managers to derive suitable market outlooks and investment decisions from a huge amount of accumulated documents. In this research, to support investment decisions, we propose a new LSTM model with self-attention mechanism that can extract important sentences in analyst visit reports. Such extraction is conducted based on the sentence scoring, which is obtained as the weights in a self-attention mechanism. In our experiments for a set of 1,390 visit reports, we demonstrate that the proposed model has about 79% accuracy for extraction on average under the 5-fold cross-validation.

1. はじめに

ICT の普及とともに,あらゆる文書が電子化され,一組織 が保有する文書資産は膨大な量となっている.組織にとって重 要かつ貴重な情報が含まれているにもかかわらず,作成された 文書情報が有効活用されずに眠ったままになっていることも多 く,有効活用したくても,その量や情報の種類が膨大であり, データベースを駆使しても有益な情報を得ることが容易でない 状況である.

投資家の資金を預かり、その資産運用を代行する運用会社 も、その例外ではない.各企業にアナリストがヒアリングし、 調査・分析を行った結果を記した往訪記録が運用会社には大量 に蓄積されている.これらには企業の財務状況や将来計画な ど、個々の企業の投資価値を判断するための重要情報が詰まっ ており、それらの情報を総合的に分析して、ファンドマネー ジャーは資金の運用方法を決める.よって、アナリストが分析 し、予測した景況感は運用会社の貴重な資産であるが、大量に 蓄積された、これら往訪記録から適切な景況分析と投資判断を 導き出すことは、熟練のファンドマネージャーであっても容易 なことではない.

そこで本研究では、ファンドマネージャーやアナリストの調 査分析にかかる労力を軽減し、適切な意思決定を容易にする ことを目的として、往訪記録から重要文章を抽出するシステ ムを開発する.3節では開発するシステムについて述べるが、 このシステムでは、まずリカレントニューラルネットの一種で ある Bidirectional Long Short-Term Memory (BiLSTM)を 用いて文章単位の埋め込みベクトルが求められる.そして、そ の埋め込みベクトルに対して自己注意機構による重みづけを 行った結果が往訪記録に付与された景況感情報と一致するよ う学習し、その自己注意の重みに基づいて、重要文章が抽出さ れる.4節では、三井住友 DS アセットマネジメント株式会社



図 1: リカレントニューラルネットの基本構造

(旧,大和住銀投信投資顧問株式会社)から提供されたアナリ スト往訪記録を用いて,トピックモデルを教師ありに拡張した Supervised Latent Dirichlet Allocation (SLDA)と提案手法 との性能比較を行い,提案手法で抽出した重要文章の精度につ いて抽出事例を示して考察する.

2. 関連研究

2.1 リカレントニューラルネット

リカレントニューラルネット (以下, RNN) は音声や言語, 動画像等の時系列データを扱うために考案されたニューラル ネットである. 図1に RNN の基本構造を示す. 時刻 t におけ る単語の埋め込みベクトル w_t が与えられたとき, セル s_t に 対する出力は,前時刻のセルの値 s_{t-1} を用いて以下のように なる.

$$\boldsymbol{s}_t = f(\boldsymbol{U}\boldsymbol{w}_t + \boldsymbol{W}\boldsymbol{s}_{t-1}) \tag{1}$$

ここで, **U** は入力層と隠れ層間の重み行列, **W** は前時刻セル と現時刻セル間の重み行列を表し, *f* は活性化関数を表してい る.通常のニューラルネットワークと異なるのは Ws_{t-1} の項 であり、前時刻のセルの値を考慮するこの値によって、RNN は各時刻間の入力変化の仕方を学習することが可能となって いる.また、RNN には出力の獲得の仕方が2種類存在する. 1つは各時刻における入力の埋め込みベクトルを得る場合であ る.この場合、時刻 t におけるセル s_t からの出力 h_t は、隠 れ層-出力層間の重みを V、活性化関数を g として以下の式の ようになる.

$$\boldsymbol{h}_t = \boldsymbol{g}(\boldsymbol{V}\boldsymbol{s}_t) \tag{2}$$

この h_t は前時刻のセルの値 s_{t-1} を含んだ値であるため,入力 w_t に時系列情報を加えたものと解釈できる.もう1つは最終セルからの出力 h_S のみを用いる場合である.この場合,最終セル s_{last} からの出力 h_S は、以下の式のようになる.また、Sは最終時刻をlastとしてS = last + 1となる値である.

$$\boldsymbol{h}_S = \boldsymbol{W} \boldsymbol{s}_{last} \tag{3}$$

他の出力を用いず、 h_S のみを用いて学習を行う場合、 h_S は 高次元空間上の各入力が、どのような順番で出現したかの情報 を保持して出力された値である。例えばwを単語の埋め込み ベクトルとすると h_S には、その文章の流れが情報として入る こととなる。

2.2 Long Short-Term Memory

Long Short-Term Memory (以下 LSTM) [Hochreiter 97] は 1997 年 Hochreiter と Schmidhuber らによって提案された リカレントニューラルネットの一種である. 一般的な RNN モ デルで時系列データの長期的な時間構造を獲得するには, フィー ドバック結合を介した再帰的信号伝播を繰り返す必要がある. 誤差逆伝播法を RNN に拡張した学習アルゴリズムでは, こ の再帰的信号伝播の過程で誤差情報が消失していく, いわゆる 勾配消失が問題となる. LSTM はこの問題を解決するためモ デルに長期記憶と短期記憶に関する項を加えている. これによ り従来のモデルでは不可能であった長期依存を学習可能にして いる. 詳細については [岡谷 15] を参照されたい.

また,時系列データが一括して与えられる場合は,t = 1, 2, ..., Tと時間方向に対して順方向に与えるだけでなく, t = T, T - 1, ..., 1と逆方向に与えることで,時系列を未来から 過去に向かって帰納的に整合を取りながら特徴量化する構造を 陽に組み込むことができる.このような順方向の LSTM と逆 方向の LSTM の両方を統合したものを Bidirectional LSTM (以下 BiLSTM) といい,順方向の出力 $\overrightarrow{h_t}$ だけでなく逆方向 の出力 $\overleftarrow{h_t}$ も特徴量として使えるようになるため通常の LSTM よりも精度がよいとされる.

2.3 自己注意機構

まず,自己注意機構を導入した表現学習モデルである,Lin らの文埋め込み表現獲得法 [Lin 17] を説明する.本論文で提 案する文書極性判定手法は,次の2つのパートからなる.ま ずはじめに,単語ごの埋め込みベクトル w を BiLSTM への 入力とし,単語ごとに時系列を考慮した埋め込みベクトルを 得る.次に,単語ごとに出力された埋め込みベクトルから自己 注意の値を求め,それをもとに作成した文書ベクトルが SVM などのクラス分類手法で分類できるよう学習させる.図2に モデルを示し,以降で詳細を説明する.

Word2Vec [Mikolov 13] などを用いて得た単語ごの埋め込 みベクトル w_t ($t = 1, 2, \dots, n$)を文章ごとにまとめたものを S とする. この S の次元は, Word2Vec で出力される単語の 次元数を v として $v \times n$ で表される.

$$\boldsymbol{S} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n) \tag{4}$$



図 2: 単語の埋め込みベクトルに対する自己注意を用いた文書 ベクトル生成モデル例

ここで, n はその文章に含まれる単語の数を表す. S を入力と して BiLSTM から出力される t 番目の単語の埋め込みベクト ルは, 順方向出力を $\overline{h_t}$, 逆方向出力を $\overline{h_t}$ として以下のように 書ける.

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(\boldsymbol{w_t}, \overrightarrow{h_{t-1}}) \tag{5}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(w_t, \overleftarrow{h_{t-1}}) \tag{6}$$

各方向の LSTM によって得た埋め込みベクトルの次元数をuとする. 2つを合わせて次元数 2uのベクトル h_s とし,これらをまとめてベクトル集合 Hとする.この Hの次元は各方向の LSTM からの出力の次元数をuとすると $n \times 2u$ で表せる.

$$\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n) \tag{7}$$

この出力ベクトル H に対して,どの単語埋め込み表現が景 況感に対するラベル予想に貢献するかを即時的に求めるため, 自己注意機構を導入する.各単語の出力ベクトル H に対する 自己注意機構による重み行列 A は,以下で与えられる.

$$\boldsymbol{A} = softmax(\boldsymbol{W}_{2} tanh(\boldsymbol{W}_{1} \cdot \boldsymbol{H}^{\mathrm{T}}))$$
(8)

 W_1 の次元数は、dを中間層の次元を表す定数として $d \times 2u$ として表せ, さらに選択的注意を与える出力数を r とおくと, W_2 の次元数は $r \times d$ で表される.この自己注意の重み係数 は多層パーセプトロンの結合荷重と似ており、これによりベ クトルh に重みの乗算を1度行っただけでは表現しきれない NOR の論理演算を、再度重みを掛け合わせることによって達 成できる. また, W_1 を乗算した後に tanh を乗算すること によって, 分離境界面から大きく離れた影響を無視し, 最後に softmax 関数を通すことによって、各入力に対する重要度を 確率で表すことが可能となる.また,注意を与える出力数 r が 1の場合、クラスを分類するのに最適な単語一つを選び、その 単語の重みのみを上げれば分類精度を上げることができる. そ のため注意の大半が1つの単語に偏ってしまうことがある。例 えば、ポジティブな文章かどうかを判定したければ、「ポジティ ブ」という単語の自己注意の値を1にすれば、それで達成さ れる. 単語は、まず Word2Vec 等で埋め込みベクトル化され ているため,影響はいくらか緩和されるが,それでも注意機構 を導入することで判定結果が偏ることがある.これを避けるた め、あるいは複数の注意を獲得する必要がある場合は、この値 を必要な注意の数だけ用意することで対応できる.これらの操作によって得た BiLSTM からの出力ベクトル集合 H,およびその注意機構の重み行列 A を用いて,文書の埋み込み行列 M は自己注意を重みとして各ベクトルを足し合わせたものとする.

$$\boldsymbol{M} = \boldsymbol{A} \cdot \boldsymbol{H} \tag{9}$$

この *M* の次元数は *r* × 2*u* で表せる.またこの手法によって 得られた自己注意の値は「その単語がどれほどラベルに影響を 与えたか」を表し,言い換えればこの手法によって「ラベルを 推定する上で重要だと判断された単語」が推定できる.

自己注意機構付き LSTM による重要文 抽出

3.1 アナリスト往訪記録

本研究では、アナリスト往訪記録から重要文章を抽出するこ とを目的としている. 往訪記録は, 各アナリストが企業へ取材 に赴き、その調査結果について「日付」「会議種別」「証券コー ド|「会社名|「担当者|「コメント」などをまとめたものであ り、どこに投資するかを決定する際には、特に「コメント」が 重要となる.「コメント」には、規則としてはじめに「担当ア ナリストは投資に対して前向きかどうか」を「ポジティブ」, 「ややポジティブ」,「ニュートラル」,「ややネガティブ」,「ネ ガティブ」の5段階評価で書き記すことになっている.そし て,その後の項目では,各企業の投資価値を判断するため,ポ ジティブとネガティブな側面を述べ、景況予想や期待感なども 交えつつ分析結果を記述する. このコメントを構成する単語数 は短いときで 4,5 単語の結論のみであり,長いときで 110 単 語前後である. これは論文やニュース記事などと比べれば非常 に短いがツイートなどに比べれば長い. また, 今回の研究の目 標として重要文章抽出を挙げているが、「ニュートラル」の評 価が与えられた文章はアナリストらもどこが重要なのかを明確 にさせていないため,訓練データには入れないことにした.

3.2 提案モデル

ここでは既存手法を改良して金融文書から重要な文章の抽出 を行うモデルを提案する.図 3.2 にその概形を示す.まず, n個 の単語からなる文章について,既存手法と同じく wrod2vec を用 いて各単語の埋め込みベクトル集合 $W_s = (w_1, w_2, ..., w_{n_s})$ を得る.なお, W_s 及び n_s の添え字 s は、ある文書におけ る文章数を S として s = 1, 2, ..., S である.このベクトルを BiLSTM の入力とし、文章ごとの埋め込みベクトル $\overrightarrow{h_s}$ と $\overleftarrow{h_s}$ を得る.この $\overrightarrow{h_s}$ と $\overleftarrow{h_s}$ は各方向の LSTM に W_s を入力した 際の出力であるが、ここで注意して欲しいのは単語ごとのベク トルを出力しているのではなく、文章ごとにベクトルを出力し ている点である.つまり、各文章における LSTM の最終出力 を用いる.

$$\overrightarrow{h_s} = \overrightarrow{LSTM}(W_s) \tag{10}$$

$$\overleftarrow{h_s} = \overleftarrow{LSTM}(W_s) \tag{11}$$

各方向の LSTM によって得た埋め込みベクトルの次元数を uとする. 2つを合わせて次元数 2uのベクトル h_s とし,こ れらをまとめてベクトル集合 Hとする.また,Hの次元は $N_S \times 2u$ であり, N_S は文書に含まれる文章の数をあらわす.

$$H = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_{N_S}) \tag{12}$$

ベクトル H を入力として文書の注意ベクトル A を得る.また本研究ではコメント内に重要文章は1つのみあると仮定し

注意の数 r は 1 としているため, その次元数は N_S である.

$$\mathbf{A} = softmax(\boldsymbol{W}_{2} tanh(\boldsymbol{W}_{1} \cdot \boldsymbol{H}^{\mathrm{T}}))$$
(13)

この式が既存手法と異なる点は入力の次元数であり,既存手法 は「文書内の単語数 $n_s \times$ BiLSTM の出力 2u」だったが提案 手法では「文書内の文章量 $N_S \times$ BiLSTM の出力 2u」である. そのため自己注意の次元数も変化し,既存手法の自己注意 Aの次元数は「注意数 $r \times$ 文書内の単語数 n_s 」だったが提案手 法では「注意数 $1 \times$ 文書内の文章数 N_S 」となっている.これ により既存手法と同様以下の文書ベクトル M を得る.

$$\boldsymbol{M} = \boldsymbol{A} \cdot \boldsymbol{H} \tag{14}$$

Aは次元数 N_S の行ベクトル, Hは次元数 $N_S \times 2u$ の行列 であるため, M は次元数 2u のベクトルとなる.最後に文書 ベクトルに重み W_3 を掛けて総和をとり, softmax 関数に通 したものを,各ラベルに割り当てられる確率 lとする.本研究 ではポジティブとネガティブの 2 クラスを想定しているため, 出力層のセル数は 2 としている.

$$\boldsymbol{l} = softmax(\boldsymbol{W}_3 \cdot \boldsymbol{M}) \tag{15}$$

4. 重要文抽出の性能評価

4.1 実験設定

性能評価には,実際のアナリスト往訪記録1,390 文書を用いた.これを提案手法で使用可能なデータとするため,以下の変換操作を行った.

- 各コメントについて先頭12文字をみてポジティブ,あるいはそれに類する単語が記されていれば1を,ネガティブ,あるいはそれに類する単語が記されていれば0をコメントの正解ラベルとして付与し,そのラベルに相当する単語は訓練データから除く.
- 2. 形態素解析には MeCab を用い, 助動詞, 助詞, 副詞, 記 号以外の単語を抽出する.
- 3. 形態素解析で得られた単語系列と正解ラベルを合わせて, 一つの文書データとする.

学習に使う損失関数には,正解ラベルと推定結果で定義される,次式のクロスエントロピー *L*を導入した.

$$L = -\frac{1}{D} \sum_{d} \log l_d \tag{16}$$

ここで, $d \in D$ は文書番号, l_d は出力の正解ラベル成分である.式(16)のLを最小化するよう誤差逆伝播法で重みを学習する.

訓練データに用いたアナリスト往訪記録 1,390 件のうち 284 件については,アナリストやファンドマネージャーによる重要 文判定が行われており,この判定結果との一致度合を評価指標 として採用した.また,もう1つの評価指標としてラベルの推 定精度を採用した.1,390 文書に対して,訓練データとテスト データが 4:1 となるように分割し,5 重交差検定を行った景況 感予測のテスト精度の平均と分散で評価する.なお,性能比較 のモデルとして,トピックモデルを教師ありの学習へと拡張し た SLDA[Mcauliffe 08] を選んだ.SLDA では,トピック分布 と単語分布を学習させた後に文章ごとのラベル推定値を求め, その中で最も推定値が正解ラベルに近かった文章を最重要文と した.



図 3: 文章ごとの埋め込みベクトルに対して自己注意機構を適用した重要文章抽出モデル.本研究では、文章ごとの自己注意 $A_i(i=1,2,\cdots,S)$ として獲得される.

表	1:	最重要文-	ー致数と	:景況感予測精度の比較	3
---	----	-------	------	-------------	---

	最重要文一致数	景況感予測精度
提案モデル	100/284	0.79 ± 0.3
SLDA	84/284	0.73 ± 0.4

4.2 評価結果と考察

アナリスト往訪記録のうち 284 件については,アナリスト やファンドマネージャーによる最重要文の判定がなされてお り,これに基づいて最重要文の一致数を求めた.結果を表1に 示す.

表1より,提案モデルがSLDAよりも優れた景況感予測が 行えており,重要文一致数から,より正確に最重要文を抽出で きていることがわかる.コメント文に対するスコアリング結果 の例を右に示す.各文章の最初に付している数値は,自己注意 機構の重み係数であり,これが高いほど重要と判定されたこと を意味する.なお,下線はアナリストが重要だと判定した文章 である.

この結果より,自己注意機構の重み係数が大きい文は,ポジ ティブな往訪記録における前向きなコメントと一致しており, ネガティブな往訪記録に対しては,ネガティブ要因を説明する 文章と一致していることがわかる.しかしながら,アナリスト が最重要と判定したポジティブとネガティブなコメントとは一 致しておらず,この二例に対しては,専門家が要求する重要度 を十分に表現できてないと言える.表1からわかるように,現 時点における重要文一致率は約35%であり,まだ改善の余地 がある.

5. おわりに

本研究では、BiLSTM で得られた文章組み込みベクトルに 対して自己注意機構を導入し、アナリスト往訪記録における重 要文章の抽出を試みた.アナリスト往訪記録 1,390 文書を用い て性能評価を行った結果、5 重交差検定による提案手法のテス ト性能は約 79% となり、専門家による重要文章判定の一致度 合についても、SLDA に比べて約 1.2 倍高い性能を得た.

抽出された重要文章は,著者が判定した景況感におおよそ 合致していたが,専門家の考える重要文章とは異なるものも少 なくなかった.これは,提案手法で獲得された文章単位の組み 込みベクトルと景況感ラベル情報の相関関係では,アナリスト 視点の重要度を十分に反映できないことを示しており,今後, アナリスト視点の重要度を組み込んだ機械学習モデルの開発が 必須と考えられる,

ややポジティブ。

(0.006)<u>A</u>商品はシェアアップが明確。
(0.96)B商品の利益も予想以上。
(0.004) ヘッドは軟調だが、C社で17年以降巻き返し、 D社買収効果も見込めそう。
(0.02)残る問題事業はマグネットのみとなった。
(0.003)16年度はMLCCでF商品巻き返しも。

ややネガ。

(0.13) A 国での B 社、C 社向けの部品売上高が想定を 下回り、D 地域の収益が期待値を下回っている。
(0.64) オートマティック車向け部品は B 社からの設計変 更が来ており、収益化が遅れそう。
(0.009) 当初は 18/3 期に収益化が見込まれていたが、今 回のミーティングでは 19/3 期までずれ込みそうなコメント。
(0.003) ただし、全社 OP に占める D 地域比率は 1 割未 満であり、むしろ 8 割を占める E 地域の収益改善の方が 重要である。
(0.006) その E 地域に関しては、F 国は堅調も 2 輪市場 の成長は鈍化傾向にある。
(0.20) 一方、G 国 2 輪向けの事業拡大への期待値は高い。

参考文献

- [Hochreiter 97] Hochreiter, Sepp, and Jrgen Schmidhuber.:Long short-term memory, Neural computation 9.8 (1997)
- [Lin 17] Lin, Zhouhan, et al.:A structured self-attentive sentence embedding. arXiv preprint arXiv.(2017)
- [Mcauliffe 08] Mcauliffe, Jon D.; Blei, David M.:Supervised topic models. In: Advances in neural information processing systems.(2008)
- [Mikolov 13] Mikolov, Tomas, et al.:Efficient estimation of word representations in vector space. arXiv preprint arXiv.(2013)
- [岡谷 15] 岡谷貴之:機械学習プロフェッショナルシリーズ・深 層学習.(2015)