業績要因の極性付与を目的とした文の区切り位置推定

高野 海斗
成蹊大学
dd186201@cc.seikei.ac.jp

酒井 浩之
成蹊大学
h-sakai@st.seikei.ac.jp

中川 慧
野村アセットマネジメント株式会社
k-nakagawa@nomura-am.co.jp

1 はじめに

近年，機械学習などの手法が注目を集め，様々な分野への応用研究が活発に行われている．金融業界でも人工知能分野の手法や技術を金融市場における様々な場面に応用することが期待されており，膨大な金融情報を分析し投資判断を支援する技術にも注目が集まっている [1, 2, 3].

一般に金融テキストには結果（事実）と評価が混在している。例えば，「今期の EPS[1)] は 100 円であり，経営努力が実った.」と「今期の EPS は 100 円であり，経営陣の見通しの甘さが露呈した.」というテキストには，今期の EPS は 100 円であるという同一の結果を両方とも伝えているものの，評価は正反対の内容を伝えている．このように，金融では同一の結果に対してどのような評価，すなわち極性を付与するかという点に市場あるいは投資家のセンチメントが含まれる．また結果を表す情報は財務データあるいは経済データが多く，定量的に取得・分析できるのに対して，評価情報はテキストにのみ存在する情報である．このように，金融テキストの極性付与は投資意思決定において重要なタスクである．

しかしながら，既存の金融テキストへの極性付与は，極性語とその極性値が組となった極性辞書を用いたものが多い [4]．極性辞書を用いた文への極性付与の問題点としては，文脈を考慮せずに極性値によるスコア付けが行われる点である．例えば，「コロナ感染者数の増加 (2.312) に伴い、売上高は減少 (-1.998) いたしました。」のような文の極性は，0.314（ポジティブ）となるが，明らかにネガティブな文である．例に挙げたように，何が増加したかによって，「増加」のスコアを変化させるべきだが，一般的な極性辞書は，そのような問題に対処できない．解決策として，「感染者数の増加 (-1.315)」，「売上の増加 (3.312)」のように表現を拡張した辞書を作成することが挙げられる [5]．しかし，全ての表現を辞書で作成することは困難であり，類似表現に対しても，辞書に登録されていなければ極性を付与できない問題がある．また，表現を拡張できたとしても，「見込んでいた売上減少はなかったものの、」のような否定表現などに対応できないことも問題点として挙げられる．

したがって，これらの問題に対処するために，文脈を考慮した上で文に対して極性付与を行うことが必要である．しかし，極性辞書を用いた文の極性付与に対して，深層学習モデルなどを用いて文に極性付与を行う場合には，解釈性の問題が残る [6].

そこで，入力文全体に対して極性を付与するのではなく，ポジティブなことが記述されている部分，ネガティブなことが記述されている部分，結果について記載されている部分に分割し，その上で分割した各部分に対して極性付与を行うことが可能なモデルの作成を研究目標とする．文を分割し，極性付与を行うイメージは以下の通りである．

最終的な研究目標

(+2.5) 収益合計は、投資信託ビジネスや投資顧問ビジネスでの資金流入が運用資産残高の拡大に寄与し、(+0.5) 引き続きビジネスは堅調だったものの、(-25.0) アメリカン・センチュリー・インベストメンツ関連損益が収益を押し下げました。

短く分割した各部分に対して，文脈を考慮し極性を付与することで，文全体に極性付与を行うよりも，解釈性の高い出力を得ることが可能となる．

1) 1 株当たりの利益

そのための第一段階として，本研究では，文を分割することが可能であるかの検討を行った．具体的には，高野らの研究 [7] で抽出した業績要因文[2] に対して，人手で文を分割するための正解 Tag をアノテーションし（111 文），事前学習済みモデルである BERT と BiLSTM モデルを組み合わせたモデルの学習を行い，このモデルを用いることで分割が可能であることを示した．

2 文の区切り位置と定義

本研究では，業績要因文をポジティブな要因が記述されている部分，ネガティブな要因が記述されている部分，結果について記載されている部分に分けることが目的である．まず，本研究で扱う業績要因文は以下のような文を指す．

業績要因文の例

広州においては、年間を通して生産活動に変動はあったものの、全体としては生産活動が好調に推移したことから、売上高は前年同期を上回る状況で推移いたしましたが、償却費及び労務費などの固定費の負担が増加したことにより利益は前年同期を下回る状況で推移いたしました。

本研究の目的は，上記のようなテキストを入力とし，以下に示すような分割を行うことが可能なモデルを学習させることである．

業績要因文の分割例

[B_Cause_Nega] 広州においては、年間を通して生産活動に変動はあったものの、[E_Cause_Nega] [B_Cause_Posi] 全体としては生産活動が好調に推移したことから、[E_Cause_Posi] [B_Result] 売上高は前年同期を上回る状況で推移いたしましたが、[E_Result] [B_Cause_Nega] 償却費及び労務費などの固定費の負担が増加したことにより [E_Cause_Nega] [B_Result] 利益は前年同期を下回る状況で推移いたしました。[E_Result]

文の区切り位置は，語尾を変更し，接続詞などの表現を用いることで，文をつなぎ直すことが可能になる位置とする．例えば，上記の文であれば，「〜変動はあったものの、全体としては〜」は，「〜変動があった。」，「しかし，全体としては〜」と変更することが可能であるため，区切り位置となっており，「〜が増加したことにより利益は〜」は，「〜が増加した。」，「その結果、利益は〜」と変更することが可能であるため，区切り位置となっている。また，本研究では分割の際に，文の分割した部分が，1.「ポジティブな要因が記述されている」，2.「ネガティブな要因が記述されている」，3.「結果について記載されている」のいずれかに分類も行う．

本研究では，学習データを人手にてアノテーションすることで作成する．そのために，まず学習データとなる業績要因文に対して Tokenizer を使用し，Token に分割する[3]．Token に分割した学習データに対して，人手でアノテーションした例を付録の表 3 に示す．

次に，本研究で使用する Tag について説明する．本研究で使用する Tag は，以下の 9 つである．

CLS: 文の開始位置
SEP: 文の終了位置
B_Cause_Posi: ポジティブ要因の開始位置
E_Cause_Posi: ポジティブ要因の終了位置
B_Cause_Nega: ネガティブ要因の開始位置
E_Cause_Nega: ネガティブ要因の終了位置
B_Result: 結果についての記載の開始位置
E_Result: 結果についての記載の終了位置
Other: 上記以外

これらの Tag を用いることで，分割位置を推定し，分類も同時に行う．

結果についての記載に関して，ポジティブとネガティブを付与しなかったのは，「売上高は〇〇百万円でした。」のような記述において，企業の規模によって同じ金額でもポジティブ・ネガティブが異なり，判断が難しいためである．

3 関連研究

文を分割するための研究は，文の自動要約に関する分野で研究が行われている．例えば，大野らは，学生のレポートが読みにくくなっている原因の一つに，修飾節が連なることによる文の長文化を挙げており，この長文を機械的に短文に変換する研究を行っている [9]．この研究では，文の修飾節として補足節・連体節・副詞節を自動的に判定する方法を

2） 企業を分析するにあたり，企業が公開している金融テキストに含まれる業績要因についての記載は，特に重要な情報であり，これらの文を本研究では業績要因文と呼ぶ [8].

3） 本研究では，東北大学の乾研究室が公開している学習済み BERT を使用するため，それに合わせた Tokenizer を使用した．

提案している．また，金融テキストと同様に，長文が多く出現するテキストとしてニュース原稿がある．ニュース原稿は，そのまま字幕にすると長すぎ場合が多い．そこで，自動的に長文を複数の短文に分割する研究が行われている [10, 11]．

これらの文分割の先行研究と本研究の大きな違いは，品詞情報を利用していない点である．入力となる文が業績要因文に限定されていることや，大量の教師なしデータを用いた事前学習済みモデルにより，同じような文脈で出現する Token に対しては似たような分散表現が与えられている可能性が高く，人手にて少量の学習データを作成することで，品詞情報を考慮しなくても分割が可能である．

また，本研究はセンチメント分析の一環でもある [12]．極性付与を行う方法の一つには，1 章でも述べたように極性辞書を用いた方法があるが，広く利用されている一般的な極性辞書（Harvard-IV-4 TagNeg）に含まれるネガティブ表現が，金融の分野において，約 4 分の 3 も該当しないことを示した研究もある [13]．したがって，センチメント分析を行うためには，その目的に特化した極性辞書や極性付与モデルを使用する必要がある．多くの金融極性辞書は，市場分析を目的に作成されているが，本研究では，企業の業績状況の分析に特化した極性付与を可能にすることを最終目標としている．

4 使用するデータ

本研究で扱う業績要因文は，高野らの手法 [7] を用いて，有価証券報告書から抽出したものを使用する．この業績要因文には，事業セグメントと業績結果文[4)]が紐づいている．本研究では，人手で作成可能な少量の学習データで学習を行う必要がある．そこで，学習データを少しでも多様性のあるものにするために，アノテーションを行う業績要因文の事業セグメントが一致しないように，ランダムサンプリングを行うことで，似たような事業に関する業績要因文が学習データに含まれないようにする工夫をしている．

ランダムサンプリングした業績要因文に対して，アノテーションを行い，111 文の学習データを作成した．アノテーションは，付録の表 3 に示したように，分割部分の先頭の Token と最後の Token に対して，Tag 付けを行った．

また，学習データで用意したデータとは別に，評価を行うためのテストデータを人手にて，20 文作成した．学習データと似たような業績要因文になることを避けるために，学習データに使用した業績要因文とは，付与されている事業セグメントが異なる業績要因文を対象にした．

4） 売上高，利益，それらの前年度比などが記載された文．

5 使用するモデル

本研究では，人手で学習データを作成するため，大量のデータを生成できない．そこで，少量のデータを Fine Tuning することで良好な結果が得られることが報告されている事前学習済み BERT を利用する [14]．事前学習済み BERT は，東北大学の乾研究室が公開しているものを使用した[5)]．

BERT は，Transformer の Encoder 部分を重ねたモデルであり，Transformer 層が全 12 層ある．本研究では，入力層から前半 6 層までのパラメータを固定した上で，後半 7 層から 12 層のパラメータを Fine Tuning することにした．

5.1 Simple BERT Model

事前学習済み BERT に対して，Token に分割した業績要因文を入力し，BERT の最終層からの各 Token の出力 768 次元に対して，線形変換を行うことで，Tag の推定を行うシンプルなモデルを Simple BERT Model とする．モデルのイメージ図を付録の図 1 に示す．

5.2 BERT BiLSTM Model

入力文に対して，Tag 付けを行うことで固有表現抽出を行う研究では，回帰型の深層学習モデルである LSTM などが有効であることが多い [15, 16]．そこで本研究では，事前学習済み BERT に対して，Token に分割した業績要因文を入力し，BERT の最終層からの各 Token の出力を，BiLSTM モデル[6)]に入力し，その出力に対して線形変換を行うことで，Tag の予測を行うモデルを BERT BiLSTM Model とする[7)]．モデルのイメージ図を付録の図 2 に示す．

6 モデルの学習

モデルの学習は，K 分割交差検証法（K-Fold Cross Validation）で行った．本研究では $K = 10$ とし，学習データをランダムに 10 分割し，9 割を訓練デー

5） https://github.com/cl-tohoku/bert-japanese
6） BiLSTM は，単方向の LSTM を双方向に拡張したモデルである [17, 18]．
7） BiLSTM の隠れ層の次元数は 512 次元，layer 数は 1 と設定．

表 1　K 分割交差検証の結果

モデル	エポック数	Macro-F1	Micro-F1
Simple BERT	66	0.957	0.989
BERT BiLSTM	71	**0.977**	0.996

表 2　テストデータを用いた評価結果

モデル	Macro-F1	Micro-F1
Simple BERT	0.863	0.968
BERT BiLSTM	**0.924**	0.983

タ，1 割を検証データとしてモデルを学習させた．

エポック数は 100 回とし，モデルの評価指標は，Macro-F1 と Micro-F1 を計算し，どのエポックのモデルを保存するかは，Macro-F1 が一番高いものを選択する[8]．Macro-F1 を評価指標に使用した理由は，「Other」の Tag 割合が他の Tag に比べて非常に多いため，全ての予測を「Other」にすることで，スコアが高くなってしまう評価指標である Accuracy よりも，良好なモデルを選択できる可能性が高いからである．また，Micro-F1 は，Tag の割合が大きい Tag の影響を受けやすいため，Macro-F1 を評価指標に採用した．

loss を計算するための損失関数は，多クラス分類であるためクロスエントロピーを使用した．しかし，前述した通り，「Other」の Tag 割合が他の Tag に比べて非常に多い不均衡なデータであるため，loss の計算では，学習データに付与された Tag の出現頻度によって重み付けを行った．

最適化関数は Adam を使用し，lr（学習率）以外はデフォルトのまま使用した．lr の値は，出力層に近い層を大きく設定し，入力層に近い層は小さく設定した．これは，学習済みモデルである BERT の Transformer 層のパラメータを微調整しつつ，ランダムに初期値を決めている最終層の線形変換や，BiLSTM 層のパラメータを重点的に学習させるための工夫である．詳細な lr の値については，付録の表 4 に示す．

モデル学習時の検証データによる各モデルの評価を平均した結果を表 1 に示す．

7　評価

学習したモデルの性能を評価するために，人手で作成したテストデータを用いて評価を行った．学習データを 10 分割して，交差検証を行っていることから，同じ数のモデルが存在するため，各モデルに対して，テストデータを入力し，各モデルの出力結果の多数決で Token に対する Tag の決定を行った．

評価結果を表 2 に示す．また，BERT BiLSTM Model に業績要因文を入力し得られる結果を付録（A.1）に示す．

8）　Macro-F1 の値が同一の場合は，エポック数の少ないものを選択する．

8　考察

表 1 に示した交差検証の結果と表 2 に示した評価結果から，シンプルな BERT モデルよりも，BiLSTM 層を追加したモデルの方が良好な結果であり，Macro-F1 も 0.924 と良好な結果であった．

正しい Tag の推定が行えなかった業績要因文を確認したところ，付録（A.1）に示した「～労務費が増加したこと等により、」のようなネガティブな分割部分に対して，ポジティブであると誤推定していた．これは，「増加」という単語がポジティブな分割に出現しやすい傾向があることが原因である．本研究で使用した業績要因文は，業績結果文から抽出できる数値情報が紐づいているため，「増加」が含まれているが，業績が前年度から大幅に下がっている業績要因文を抽出し，学習データを増やすことで解決できないか，今後検討を行っていきたいと考えている．また，「減少」が含まれているが，業績が前年度から大幅に上がっている業績要因文など，紐づいている情報などをうまく活用して，意図的に分類が難しい学習データを 100 文ほど増やすことで，推定にどのような影響があるのかの分析も行っていきたいと考えている．

9　まとめと今後の展望

本研究によって，少量のデータを人手で作成し，事前学習済みモデルを利用したモデルを用いることで，業績要因文の分割が可能であることを示した．業績要因文は，１文でいくつかの要因について記載されているものが多く，ポジティブなこととネガティブなことの両方が記述されている文も多いが，本研究で学習したモデルを用いることで，それらを分割することが可能である．

本研究で使用した業績要因文は，高野らの研究 [7] で抽出した業績結果文が紐づいたテキストデータであるため，今後は，これらの業績結果文に含まれる数値情報を目的変数とし，分割したポジティブな記述とネガティブな記述を入力とすることで，入力テキストに対して，極性付与を行うモデルの検討を行っていく予定である．

参考文献

[1] 和泉潔, 松井藤五郎. 金融市場における最新情報技術：8. 金融テキストマイニング研究の紹介. 情報処理, Vol. 53, No. 9, pp. 932–937, 2012.

[2] 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.

[3] 和泉潔, 坂地泰紀, 伊藤友貴, 伊藤諒. 金融テキストマイニングの最新技術動向 (特集 ai の金融応用 (実践編)). 証券アナリストジャーナル, Vol. 55, No. 10, pp. 28–36, 2017.

[4] Tomoki Ito, Hiroki Sakaji, Kota Tsubouchi, Kiyoshi Izumi, and Tatsuo Yamashita. Text-visualizing neural network model: understanding online financial textual data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 247–259, 2018.

[5] 今井康太, 酒井浩之, 高野海斗, 北島良三, 末廣徹, 稲垣真太郎, 木村柚里. 債券市場における金融極性辞書の自動構築. 第 25 回金融情報学研究会, pp. 38–43, 2020.

[6] David Gunning and David William Aha. DARPA's explainable artificial intelligence program. *AI Magazine*, Vol. 40, No. 2, pp. 44–58, 2019.

[7] 高野海斗, 酒井浩之, 北島良三. 有価証券報告書からの事業セグメント付与された業績要因文・業績結果文の抽出. 人工知能学会論文誌, Vol. 34, No. 5, pp. 1–22, 2019.

[8] 酒井浩之, 松下和暉, 北島良三. 学習データの自動生成による決算短信からの業績要因文の抽出. 日本知能情報ファジィ学会誌, Vol. 31, No. 2, pp. 653–661, 2019.

[9] 博之大野, 宏誠稲積. 修飾節に着目した長文における短文化への試み. Technical Report 12, 東京医療保健大学, 青山学院大学, 2020.

[10] 福島孝博, 江原暉将, 白井克彦. 短文分割の自動要約への効果. 自然言語処理, Vol. 6, No. 6, pp. 131–147, 1999.

[11] 江原暉将, 福島孝博, 和田裕二, 白井克彦. 聴覚障害者向け字幕放送のためのニュース文自動短文分割. Technical Report 65(2000-NL-138), 通信・放送機構／NHK, 通信・放送機構／追手門学院大学, 通信・放送機構, 通信・放送機構／早稲田大学, 2000.

[12] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, Vol. 89, pp. 14–46, 2015.

[13] Tim Loughran and Bill Mcdonald. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, Vol. 66, No. 1, pp. 35–65, 2011.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[15] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681, 1997.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[17] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Vol. 4, pp. 2047–2052, 2005.

[18] Alex Graves, Navdeep Jaitly, and Abdel rahman Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

A　付録

A.1　テストデータの出力結果

分割に成功した文

・[B_Cause_Posi] 電線事業につきましては、当事業の主要な市場である建設・電販向けの売上は、公共事業・設備投資の持ち直しがみられるものの、[E_Cause_Posi] [B_Cause_Nega] 銅価格が前連結会計年度よりも大幅にダウンしている影響で [E_Cause_Nega] [B_Result] 売上高は 5,741 百万円 (前年同期比 9.3%減) と減少しました。[E_Result]

・[B_Cause_Posi] 電子デバイスにつきましては、マイコンは堅調に推移しましたが、[E_Cause_Posi] [B_Cause_Nega] パワー半導体が大きく減少しました。[E_Cause_Nega]

分割に成功したものの分類に失敗した文

・[B_Cause_Posi] 利益面につきましては、人員不足による人材派遣の利用により労務費が増加したこと等により、[E_Cause_Posi] [B_Result] 営業損失 129 百万円 (前年同期は営業損失 89 百万円) となりました。[E_Result]

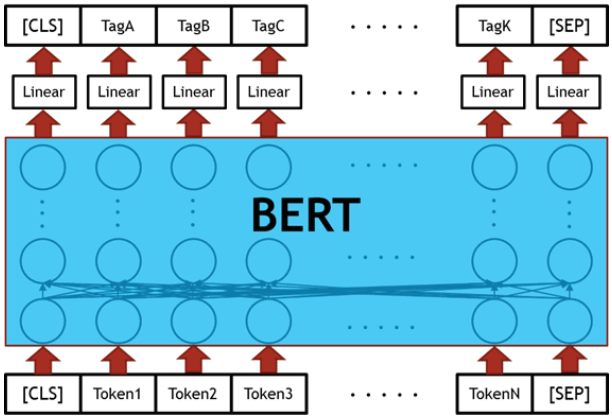

図 1　Simple BERT Model のイメージ図

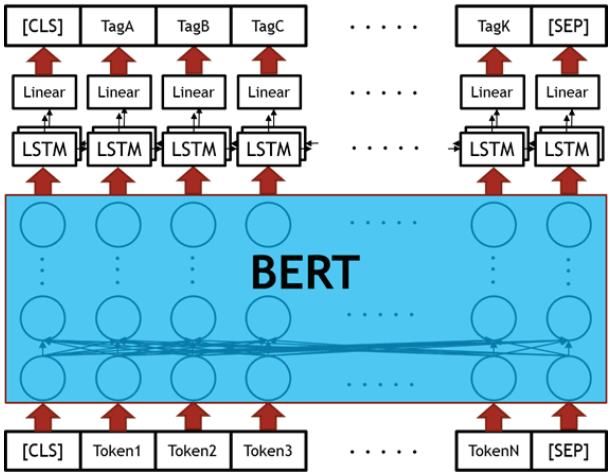

図 2　BERT BiLSTM Model のイメージ図

表 3　学習データの例

Token	Tag
[CLS]	[CLS]
広州	[B_Cause_Nega]
において	[Other]
は	[Other]
、	[Other]
年間	[Other]
を通して	[Other]
生産	[Other]
活動	[Other]
に	[Other]
変動	[Other]
は	[Other]
あっ	[Other]
た	[Other]
ものの	[Other]
、	[E_Cause_Nega]
全体	[B_Cause_Posi]
として	[Other]
は	[Other]
：	：

表 4　モデルの各層の学習率

layer	lr
BERT Layer7	5e-10
BERT Layer8	5e-9
BERT Layer9	5e-8
BERT Layer10	5e-7
BERT Layer11	5e-6
BERT Layer12	5e-5
BiLSTM Layer	1e-4
Linear Layer	1e-4