F-018

経済テキストをベクトル化するための Sentence-BERT を用いたファインチューニング

EP20101 丸山大輝　　指導教授: 松井藤五郎

1　はじめに

近年では、テキスト分析を用いて経済市場の動きを予測する研究が盛んに行われている。中でも、2018 年に Google が発表した BERT [1] という自然言語処理モデルが、当時のモデルと比べてさまざまなタスクにおいて最高スコアを叩き出している。

本研究では、その BERT から得られる分散表現をファインチューニングによってさらに精度の高い分散表現を出力できるようにし、精度の高い分散表現を株価を予測する指標とすることを目的としている。

2　対象とする問題

2006 年の日本経済新聞から選択した経済に関する記事 20 件を図 1 に示す。label はその記事がポジティブ（1）かネガティブ（0）かを表しており、Chat-GPT を用いてラベリングを行った。これを BERT に入力し、出力された 768 次元の分散表現を UMAP [2] を用いて 2 次元に削減した結果を図 2 に示す。左上部分とそれ以外で分かれているが、縦軸の幅が大きいのでそれぞれの分散表現は似たようなものであるとは言い難い。

また、本研究では、日本経済新聞を用いて Sentence-BERT によるファインチューニングすることを検討している。しかし、その際に用いる triplet データの作成には何かしらのラベルが必要であるが、日本経済新聞にはそのようなラベルは存在しない。故に、日本経済新聞を用いてのファインチューニングが行えないという問題がある。

本文	label
【ニューヨーク＝発田真人】二十八日の米国株は大幅反落した。債券は四日ぶりに反発し、ドルは売	0
昨年一年間の中国の名目国内総生産（ＧＤＰ）は約二百六十兆円と前年比一割程度の成長を続けてい	0
三月は利付国債の大量償還が予定されている。債券市場では、投資家に戻った償還資金が再び債券投	0
東芝の株価がさえない。好業績を背景に一月十三日に八百十五円の昨年来高値を付けたが、その後は	0
コマツの二〇〇六年三月期の連結業績（米国会計基準）は、営業利益が前期比六割増の千七百億円弱	1
◇携帯電話向け情報配信のインデックスが続落。ジャスダックの売買代金上位には、直近上場銘柄を	0
◇債券相場は反落。日銀が量的緩和策解除の際に、長短金利の上昇抑制策を検討しているとの報道を	0
◇東京の粗糖は続落。為替の円高・ドル安を材料に売られた。ただ前日のニューヨーク相場は上昇し	0
◇二十八日のニューヨーク商品取引所の金は急反発。ドルが対主要通貨で下落したのを材料に金の買	1
投信シフト鮮明　地方銀行が株式や外国為替相場の動き次第で価格や利回りが変動する個人向け投資商	1
【ニューヨーク＝藤田和明】米ネット検索大手グーグルのジョージ・レイエス最高財務責任者（ＣＦ	0
一日の東京外国為替市場で円相場は上昇し、一時、一カ月ぶりの円高・ドル安水準となる一ドル＝一	1
一日の東京株式市場で日経平均株価は大幅反落した。前日の米国株式相場の急落を嫌気した外国人投	0
日本郵政公社は一日、二月末の投資信託販売残高が前月末比三七％増の九百三十四億八千百万円にな	1
◇ＪＡＬが三日ぶりに反発。株主総会後に西松遥取締役が社長兼グループＣＥＯ（最高経営責任者）	0
【シカゴ＝山下真一】二十八日の外国為替市場でカナダドルが高騰。対米ドルで一時、一カナダドル	1
円は対ユーロで２日続落。１ユーロ＝１３８円台前半で推移している。朝方は対ドルでの上昇につれ	0
一日の東京株式市場で日経平均株価は大幅反落し、五営業日ぶりに一万六〇〇〇円を下回った。前日	0
東証では日経平均株価が反落。前日の米株式相場が大幅に下落したことを受け、ほぼ全面安の展開と	0
ジャスダックは日経ジャスダック平均株価が続落。米国株安などで投資家心理が冷え込み、時価総額	0

図 1　2006 年の経済に関する記事 20 件

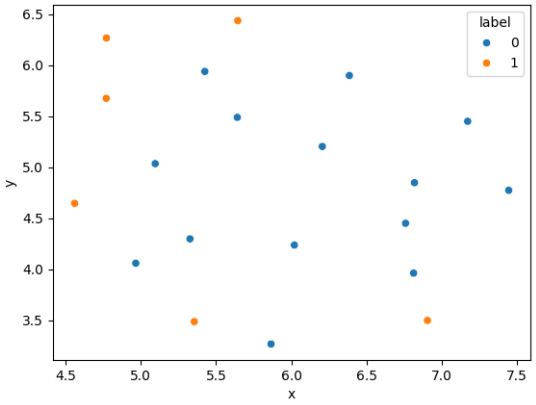

図 2　2006 年の経済に関する記事の分散表現

Algorithm 1 triplet データ作成アルゴリズム

$X \leftarrow []$
文書集合 D に含まれる文書の全組み合わせを作成し P とする
while X の長さが必要数になるまで **do**
　$(d_1, d_2) \leftarrow P$ からランダムに選択
　if $d1$ と $d2$ のラベルが同じなら **then**
　　$d_3 \leftarrow d1$ とラベルが異なる文書からランダムに選択
　　(d_1, d_2, d_3) を anchor-positive-negative として X に追加
　end if
　(d_1, d_2) を P から削除
end while

3　提案手法

提案手法の流れを図 4 に示す。chABSA-dataset [3] という上場企業の有価証券報告書 (2016 年度) をベースに作成されたデータセット（以降、chABSA と記す。）を用意する。chABSA を用いることで、ラベリングが行われていない日本経済新聞の経済に関わる記事の代用としている。rating はその文書がポジティブかネガティブかを表している。次に、東北大学によって作成された事前学習モデル [4] に対して、Sentence-BERT の目的関数を Triplet Object Function [5] としてファインチューニングを行うため、約 2800 の文書で構成されている chABSA を加工して今回は 10000 件の triplet データを作成する。

triplet データは、anchor, positive, netative の 3 つの文から成るデータであり、positive は anchor と同じラベルを持ち、negative は anchor と異なるラベルを持つ。アルゴリズム 1 に triplet データを作成する際のアルゴリズムを記す。まずは

	anchor	positive	negative
0	当連結会計年度におけるわが国経済は、企業収益や雇用・所得環境の改善が進み、緩やかな回復基調で…	当社グループの主力事業が属するインターネット広告市場は、当年度においても広告市場全体の伸びを…	その結果、売上高は1,173百万円(前連結会計年度比9.5%減)、営業利益は25百万円(同1…
1	産業別には、国内の自動車業界は、各社の新型車発売の効果により後半から販売が回復し、全体として…	消耗品である精密加工ツールは、メモリの薄化需要の高まりと顧客の高い設備稼働率に比例して、グラ…	その一方で、既存サイトの一部では会員数の減少も見られ、また、新規ファンクラブサイトの開設も想…
2	受注高は193百万円(前事業年度比1,972.9%増)、完成工事高につきましても99百万円(…	当業界においては、「大手金融機関におけるグローバル展開」や「事業領域の拡大に向けたIT投資」…	営業損益につきましては、営業利益1億45百万円となり、前期に比べ28百万円、16.6%の減少…
3	化学設備プラントおよび一般産業用プラント建設は、請負工事の増加により、前期に比べ増収となりました	建材、住宅機器等の住設建材群は、総じて市場の伸びと同程度の拡大を図ることができ、当社グループ…	運搬機械事業は、三菱重工業株式会社の子会社である三菱重工マシナリーテクノロジー株式会社の販売…
4	米国経済は雇用・所得環境の改善が下支えとなり、回復が続いており、欧州経済も英国の欧州連合離脱…	FA部門の連結売上高は、1,750億16百万円(前期比2.8%増)、全連結売上高に対する構成…	しかしながら、営業戦略の不徹底及び作業スキルのブラッシュアップ不足により、店舗オペレーション…
…	…	…	…
9995	当連結会計年度におけるわが国経済は、雇用、所得環境、企業業績の改善が続いており、景気全体とし…	需要分野別では、上期の為替影響による輸出の低調が影響した化学分野では前期比1.4%減となりま…	売上高は2,701億円で前連結会計年度比153億円(5.4%)の減収となり、営業利益は319…
9996	新造船の流入圧力はあったものの、12月にOPECで減産が合意されるまで中東各国が増産を続けた…	経常利益は、事業費が増加したものの、支払備金繰入額の減少や保険引受収益および資産運用収益の増…	当業界におきましては、人手不足を主因とする人件費関連コストの上昇並びに消費者の強い節約志向な…
9997	また、前連結会計年度における減損損失の計上により、のれん償却費等の固定費が減少しました	国内においては、こうした世界経済の影響や、個人消費が底堅く推移する中、下半期から円安・ドル高…	また、親会社株主に帰属する当期純利益につきましても、2,824億円と前連結会計年度に比べ1,…

図 3　作成した triplet データ

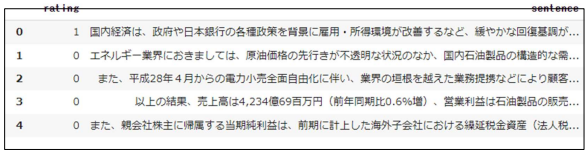

図 4　提案手法の流れ

じめに、文書集合 D に含まれる文書の全ての組み合わせを生成する。そこからランダムに anchor と positive の組を選択し、二つの文書が同じラベルであればラベルが異なる文書から negative をランダムに選択することによって triplet データを生成する。

アルゴリズム 1 で作成した triplet データを図 3 に示す。今回作成した triplet データは、anchor がポジティブであるものは 6828 件、ネガティブであるものは 3172 件で構成されている。

4　実験結果

ファインチューニングの際のハイパーパラメータは以下のように設定した。

- BATCH_SIZE = 16
- NUM_EPOCH = 1
- EVAL_STEPS = 1000
- ARMUP_STEPS = int(len(train_dataset) // BATCH_SIZE * 0.1)

図 4 の流れでファインチューニングを行い、そのモデルに図 1 の記事を入力することで有効性を確かめた。Sentence-BERT によるファインチューニングをしたモデルに図 1 を入力し、出力されたベクトルを 2 次元にした結果を図 5 に示す。図 2 と比較すると、ポジティブな記事の分散表現が 4 つ近い値の分散表現を持っていることが確認できる。

また、縦軸の幅の値が小さくなっていることも確認できる。

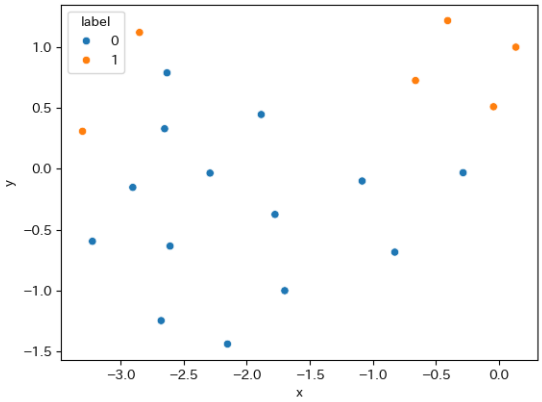

図 5　ファインチューニング後の分散表現

5　まとめと考察

本論文では、BERT はクラスタリングには向かない、日本経済新聞を用いた BERT のファインチューニングが難しいという問題に対して、日本経済新聞の経済に関する記事に代わって chABSA による Sentence-BERT を用いたファインチューニング手法を提案した。そして、chABSA を用いた Sentence-BERT によるファインチューニングを行うことで、似たような意味を持つ文書が似たような分散表現を持つようになった。この結果が得られたことで、chABSA を用いた Sentence-BERT によるファインチューニングの有効性を確かめることができた。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT 2019*, pp. 4171–4186

[2] Leland McInnes, John Healy, Nathaniel Saul, Lukas Großberger (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861

[3] KUBO Takahiro, NAKAYAMA Hiroki (2018). chABSAL: Aspect Based Sentiment Analysis dataset in Japanese. https://github.com/chakki-works/chABSA-dataset

[4] 東北大学 自然言語処理研究グループ. Pretrained Japanese BERT models released/日本語 BERT モデル公開, https://github.com/cl-tohoku/bert-japanese, 2024 年 1 月 23 日参照

[5] Nils Reimers, Iryna Gurevych (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019*:3982–3992