金融テキストからの類似文の自動収集

吉野綾音　酒井浩之　永並健吾

成蹊大学　理工学部　情報科学科

us192136@cc.seikei.ac.jp, {h-sakai, kengo-enami}@st.seikei.ac.jp

概要

本研究では有価証券報告書データから類似文を自動的に収集する手法を提案する．タグと文末表現が一致する文ペアの中で，単語の TFIDF 値を要素とした単語ベクトルから求めた cos 類似度とレーベンシュタイン距離との調和平均を算出し，類似文を収集する．収集した類似文を学習データとして Sentence-BERT で学習モデルを作成し，その学習モデルによる決算短信からの業績要因文の抽出により，収集した類似文を評価した．

1. はじめに

近年，ビッグデータの活用やテキストマイニング技術などテキストデータの有効利用が注目されている．テキストマイニングには多くのテキストデータが必要であるが，学習データ等はまだ手動で作成，収集を行なっていることも多い．特に類似文のペアは様々な言語処理に活用できる．例えば，類似文からは言い換え表現の獲得ができる．また，高精度な文ベクトルを生成する手法である Sentence-BERT[1]において，類似文のペアを学習データとして使用している．その一方で，類似文のペアの作成は人手で行っていることがほとんどであり，大量のデータを用意できないのが現状である．類似文を自動的に収集することができれば，データ作成の労力を大きく抑えられる．また，自動収集によって多くの類似文ペアのデータが入手できるため，言い換え表現の獲得や Sentence-BERT など類似文を利用する様々な技術で役立つと考える．そこで本研究では，類似した内容の多い金融テキストから，類似文のペアを自動的に収集することを目的とする．金融テキストは企業から毎年発行されるが，同じ企業が発行している金融テキストに記載される内容はほぼ同じことが多い．そのため，類似した文のペアが含まれると考えられる．本研究では，そのような金融テキストの性質を利用して，類似文のペアを自動的に収集することを目的とする．

これまでに金融テキストからの情報抽出に関してはいくつかの研究が行われている[2][3][4]．文献[2]では，決算短信から手がかり表現と企業キーワードを獲得し，業績要因文を抽出している．文献[3]では，手がかり表現を拡張して学習データを自動的に作成し，深層学習によって有価証券報告書から業績要因文を抽出している．文献[4]では，[3]と同様に学習データを自動的に作成し，深層学習によって決算短信から業績要因文を抽出している．いずれも決算短信や有価証券報告書から業績要因文を抽出する手法を提案している．これらの研究に対して，本研究では金融テキストとして有価証券報告書を対象とし，業績要因文に限らず，文書内の類似文をすべて収集する．

言い換え表現の獲得に関する研究は文献[5][6][7]がある．文献[6]では多言語パラレルコーパスを利用した言い換え表現の判定，収集であり，本研究では文全体での類似した文を収集するという点が異なる．類似した文のペアには表現の言い換えによって類似している文も含まれており，本研究で収集した類似文から言い換え表現の獲得を行うことも可能である．文献 [7]では同義語のグラフを構築し，グラフをもとにペアワイズ類似度を計算，単語単位の類似度を合計して文同士の類似度を測る手法を提案している．本研究では，cos 類似度とレーベンシュタイン距離を組み合わせて文同士の類似度を算出し，類似文を収集するという点に違いがある．

日本語の Sentence-BERT に関しては文献[8]がある．この研究ではスタンフォード NLI コーパスを日本語に翻訳し，BLEU のスコアとクラウドソーシングを利用することで類似文を収集している．本研究では金融テキストから類似文を自動的に収集する点が異なるが，収集した類似文は金融に特化しており，金融テキストを対象とした研究においては精度の向上が望めるという違いがある．

２．提案手法

本提案手法は以下の 4 つの Step で構成される.

Step 1: 有価証券報告書データから，文末表現が一致する文を抽出

Step 2: Step1 で抽出された文において，2つの文の類似度を単語の TFIDF 値を要素とする単語ベクトルによる cos 類似度を算出

Step 3: Step1 で抽出された文において，2つの文の正規化レーベンシュタイン距離を算出

Step 4: cos 類似度と正規化レーベンシュタイン距離の調和平均を算出し，調和平均が閾値より高い2つの文を類似文として収集

2. 1 金融テキスト

本研究で使用する金融テキストとして，上場企業が発行している有価証券報告書を使用する. 有価証券報告書は EDINET[1]から XBRL 形式のファイルで取得した 4593 社の有価証券報告書のテキストデータを使用する.

XBRL 形式のテキストデータには文ごとに文の内容を示すタグが付与されている. タグは全部で 479 種類であった. 例えば「BusinessRisksTextBlock」タグは，「そのような研究開発活動の停滞により、当社グループの業績が悪影響を受ける可能性があります。」のような事業リスクに関する文に付与されている. タグが同じ文は内容が近しいと考えられるため，同じ企業の同じタグを持つ文を比較する.

2. 2 文末表現による類似文候補の取得

文末文節と，その文末文節に係る文節を結合した文字列を文末表現とし，文末表現が一致する文を抽出する. 文末表現を得るための係り受け解析には Cabocha[9]を利用した. 内容が近い文ペアは文末表現が一致すると考えられるため，同一のタグをもち，かつ，文末表現が一致する2つの文を類似文候補として抽出する. しかし，これだけでは図1のように類似文として不適切な文ペアが多く抽出されるため，文間の類似度を求めることで，類似文として不適切な文ペアを除去する.

社会全般にわたる重大な品質問題など、当社グループの取り組みの範囲を超えた事象が発生した場合には、業績に影響を及ぼす可能性があります。
販売量・単価共にこの季節変動及び気候・天候条件に影響を受け易く、その変動が大きい場合は、業績に影響を及ぼす可能性があります。

図 1 類似文として不適切な文ペアの例

2. 3 類似文の判定

抽出された類似文候補の2文間の類似度を求め，類似文として不適切な文ペアを除去する. ここで，類似度は文に含まれる単語の TF・IDF 値を要素とする単語ベクトルの cos 類似度を用いる. 以下の式を用いてある企業における単語aの TF・IDF 値を求める.

$$TF(a,\ x)=\frac{\text{文}x\text{における単語}a\text{の出現頻度}}{\text{文}x\text{における全単語の出現頻度の和}}$$

$$TF\cdot IDF(a,\ x)=TF(a,\ x)\times\log\left(\frac{N}{\text{単語}a\text{を含む文の数}}\right)$$

ここで，N はある企業が発行している有価証券報告書の集合における文の総数である.

TF・IDF 値で文の単語ベクトルを生成し，単語ベクトルによる文間の cos 類似度を求める. cos 類似度が高い文ペアを類似文とすることで，不適切な文ペアを除去する.

本研究では全く同じ文や数字だけが異なる文ではなく，内容は類似していながら表現の異なる文を類似文として収集したい. しかし，タグと文末表現の一致と cos 類似度による収集では，内容が類似している文ではあるものの，表現の異なる文の収集ができない. そこで本研究では類似度の判定として，cos 類似度に加え正規化レーベンシュタイン距離を利用する. レーベンシュタイン距離とは編集距離のことで，1 文字の挿入・削除・置換で，一方の文字列をもう一方に変形するために必要な手順の最小回数を表す.

本手法では，cos 類似度と正規化レーベンシュタイン距離の調和平均が大きい文ペアを類似文として判定する. 本来，レーベンシュタイン距離が小さいとき文同士は類似していることになるが，本手法では大きいものを収集した. なぜなら，文間類似度を表す cos 類似度と，編集にかかるコストを表すレーベンシュタイン距離がとも

[1] https://disclosure.edinet-fsa.go.jp/

に大きい文ペアを抽出することで，同じ単語を異なる配置で使っている文を類似文として獲得できると考えるためである．cos 類似度と正規化レーベンシュタイン距離の調和平均の閾値は 0.5 以上とした．

本手法により 401,347 ペアの類似文を収集した．本手法による類似文として判定された文ペアの例を以下に示す．

花種子につきましては、トルコギキョウやヒマワリの売上が伸びたことなどから、前期比増収となりました。
花種子につきましては、為替の影響で欧州、南米では減収となりましたが、中国ではトルコギキョウ、北米ではヒマワリやトルコギキョウなどが好調に推移したことから、前期比増収となりました。

図 2　本手法による収集した類似文の例

3. 評価

本手法で収集した類似文を学習データとして Sentence-BERT でモデルを生成し，生成されたモデルを使用することで収集した類似文の評価を行う．適切な類似文が収集できていれば，その類似文を学習データとして用いて作成したモデルは適切な文ベクトルを生成できる．ここで，評価タスクとして決算短信からの業績要因文の抽出[2][4]を設定する．このモデルを利用して決算短信から業績要因文の抽出を行い，抽出された業績要因文の精度，再現率で評価する．

収集した類似文を用い，Sentence-BERT のモデルを生成する．なお，学習データとして利用する非類似文は，同じ企業の同じタグの文の中で最も cos 類似度が低いものを収集した．そして，文献[4]の手法にて作成した学習用の業績要因文データと，テストデータである決算短信の文との類似度を，Sentence-BERT で学習したモデルを用いて求め，学習用の業績要因文との類似度が高い文を決算短信から抽出する．

抽出された業績要因文の精度と再現率によって，収集した類似文の適切性を評価する．類似度の閾値により，精度，再現率，F 値が変化する．表1に類似度の閾値における，精度，再現率，F 値を示す．

表 1　評価結果

閾値	再現率(%)	精度(%)	F 値
0	100	38	55.1
0.1	99	39	56.0
0.2	98	42	58.8
0.3	97	43	59.6
0.4	96	45	61.3
0.5	92	53	67.3
0.6	10	79	17.8

比較手法として，以下の 2 種類の手法で収集した類似文で Sentence-BERT のモデルを学習し，比較した．

1) cos 類似度と正規化レーベンシュタイン距離の調和平均により抽出した類似文から，同一文字列が含まれる文ペアを除いた類似文

2) cos 類似度のみで抽出した類似文

cos 類似度と正規化レーベンシュタイン距離の調和平均により抽出した類似文のデータには，同一文字列が含まれる組み合わせが散見される．なるべく異なる表現の類似文を収集したいため，1) では同一文字列を含む文ペアを除き学習データとした．2) では，ベースラインとして，cos 類似度のみで類似文の判定を行なった．閾値は 1) と同様に 0.5 以上とした．表2に比較手法の精度，再現率，F 値を示す．表2では，各手法で F 値が最も大きい時の評価値を示した．閾値は本手法では 0.54 の時，1)では 0.70 の時，2)では 0.71 の時の評価値である．

表 2　提案手法との比較

	本手法	1)同一文字列を除く	2)cos 類似度のみ
再現率(%)	84	98	95
精度(%)	61	43	44
F 値	70.7	59.8	60.1

本手法と比較手法において，閾値を 0.01 ずつ変化させたときの精度，再現率のグラフを示す．

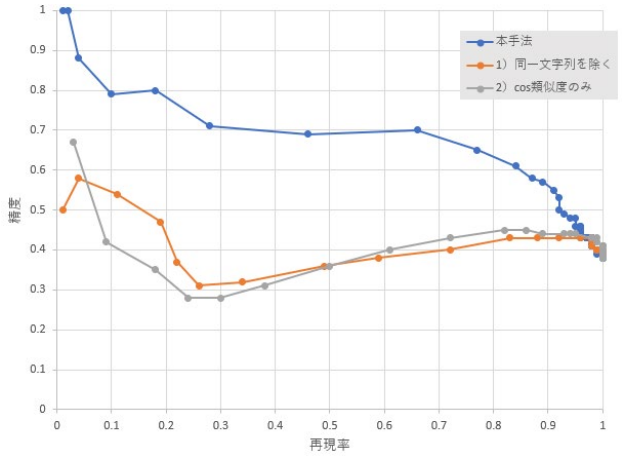

図 3　提案手法と比較手法の評価結果

4. 考察

金融テキストを利用した類似文の収集において，収集した類似文を用いて Sentence-BERT のモデルを学習した．そして，作成されたモデルによって決算短信からの業績要因文の抽出を行い，その精度と再現率で収集した類似文の評価を行った．類似度の閾値が 0.54 の時，精度が 61%，再現率が 84%，F 値が 70.7 となった．3 種類の手法によって収集した類似文ペアで学習モデルを作成したが，本手法が最も良い結果が得られ，同一文字列を除く場合は精度が落ちた．言い換えや内容の順番の入れ替えによる類似文とは言い難い同一文字列を含む文ペアを除くことは精度向上につながると考えた．しかし，正規化レーベンシュタイン距離の値を考慮しても調和平均の値が大きく，類似文として完全にノイズであるとはいえない．類似文の総数が同一文字列を含む場合は約 40 万文，含まない場合は約 38 万文となり，同一文字列を含む文ペアを除いたことによって学習データが少なくなり，それにより精度が落ちたと考える．さらに，本研究では類似文を自動的に収集しているため，データ量が膨大となり，すべての文ペアを手作業で確認ができない．そのことから，同一文字列を含む文ペアを除いた場合の学習データには，例えば図 4 のようなノイズが含まれていた．学習データが少なくなったことで，学習データにおけるノイズの影響が大きくなり，それにより精度が落ちたと考える．

服飾資材関連では，スポーツアパレルメーカー向け付属品の売上高が増加しました
生活産業資材関連では，映像機器向け付属品の売上高が増加しました

図 4　類似文として不適切な例

また，Sentence-BERT のモデルの作成の際に必要となる非類似文は，単純に最も cos 類似度が低くなる文にしたが，「あります」など極端に短い非類似文があり，そのような非類似文がノイズになったと考える．この非類似文の収集についても，含まれる文字数を考慮したり，レーベンシュタイン距離等を組み合わせたりすることで，精度の向上が目指せるのではないかと考える．

類似文の抽出では，数字だけが異なる文の組み合わせを省くため，「円」や「%」が含まれる文は類似文として取得しなかった．しかし，「名」「年」「株」「単元」なども単位として頻繁に使われている．これらの単語についてもノイズとして除くことで精度の向上に繋がる可能性がある．しかし，それによって「株式会社」や「当事業年度」などのよく使われる単語も省かれるため，金融テキストに特有の言い回しやよく使われる単語のリストを作成し，省く単語の適切な条件を設定する必要がある．

5. むすび

本研究では有価証券報告書データから類似文を収集する手法を提案した．本手法では cos 類似度と正規化レーベンシュタイン距離との調和平均を算出し，類似文を収集した．決算短信からの業績要因文の抽出による評価の結果，精度が 61%，再現率が 84%，F 値が 70.7 となり，同一文字列を含む文を除いた場合，cos 類似度のみの場合と比較して良好な結果を得ることができた．今後の課題として，単位を含む文の除去，文末，文頭表現の一致の条件を適切に設定して精度を向上させたい．

参考文献

[1] Nils Reimers, Iryna Gurevych: “Sentence-BERT: Sentence Embeddings using Sia-mese BERT-Networks”, Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3982-3992, November 3-7, 2019.

[2] 酒井 浩之, 西沢 裕子, 松並 祥吾, 坂地泰紀: “企業の決算短信 PDF からの業績要因の抽出”, 人工知能学会論文誌, Vol. 30, No. 1, pp. 172-182, 2015.

[3] 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, p.wd-A_1-22, 2019.

[4] 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.

[5] 乾 健太郎, 藤田 篤: “言い換え技術に関する研究動向”, 言語処理学会論文誌, Vol. 11, No. 5, pp. 151-198, 2004.

[6] 柏岡 秀紀: “多言語パラレルコーパスを利用した言い換え表現グループの構築と分析”, 言語処理学会論文誌, Vol. 11, No. 5, pp. 3-18, 2004.

[7] Youhyun Shin, Yeonchan Ahn, Hyuntak Kim, Sang-goo Lee: “Exploiting Synonymy to Measure Semantic Similarity of Sentences”, IMCOM 15: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, No. 40, pp. 1-4, 2015.

[8] Naoki Shibayam, Hiroyuki Shinnou: “Construction and Evaluation of Japanese Sentence-BERT Models”, In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 731-738, 2021.

[9] Taku Kudo, Yuji Matsumoto: “Fast Methods for Kernel-Based Text Analysis”, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 24-31, 2003.