テキストマイニングを利用した テーマに関連する上場企業検索ツールの開発

Development of Search Tool for Listed Company Related to Themes Using Text Mining

平野正徳 [1*] 坂地泰紀 [1] 木村笙子 [2] 和泉潔 [1] 松島裕康 [1] 長尾慎太郎 [2] 加藤惇雄 [3]
Masanori HIRANO[1] Hiroki SAKAJI[1] Shoko KIMURA[2]
Kiyoshi IZUMI[1] Hiroyasu MATSUSHIMA[1] Shintaro NAGAO[2] Atsuo KATO[3]

[1] 東京大学 大学院工学系研究科
[1] School of Engineering, The University of Tokyo
[2] 大和証券投資信託委託株式会社 調査部
[2] Research Department, Daiwa Asset Management Co. Ltd.
[3] 株式会社大和総研 フロンティアテクノロジー本部
[3] Frontier Technologies Research & Consulting Deptartment, Daiwa Institute of Research Ltd.

Abstract: We propose a scheme for selecting stocks related to a theme. This scheme was designed to support fund managers who are building themed mutual funds. Our scheme is a type of natural language processing method and based on words extracted according to their similarity to a theme using word2vec and our unique similarity based on co-occurrence in company information. We used data including investor relations and official websites as company information data. We also conducted several other experiments, including hyperparameter tuning, in our scheme.

1 はじめに

近年，日本において，個人投資家の数が増えている．個人投資家の増加に対応し，個人投資家をサポートするさまざまな技術が開発されている．ニュース記事の中から自動的に景気動向を示す表現を抜き出す手法 [Sakaji 08] や，上場企業が公開する，共通形式の決算速報である決算短信の中から自動的に業績関連文を取り出す手法 [Kitamori 17]，金融テキストのセンチメントを可視化する解釈可能なテキストマイニング [Ito 18b, Ito 18a] などがあげられる．

日本の個人投資家の間では，特に投資信託と呼ばれる，投資家から集めた資金で投資信託会社などが資金運用をして，成果を投資家に還元する金融商品が人気がある．似たような金融商品の一種として，世界中で取引されている Exchange Traded Funds (ETF) が有名であるが，投資信託と ETF は大きく異なる．ETF は一般に，それぞれの ETF を構成する構成銘柄の比率は固定されている．一方で，投資信託は，投資信託を管理している金融機関や投資信託会社などが，動的に構成銘柄やその構成比率を変更させている．

投資信託の中には，テーマ型投資信託 (テーマ型ファンド) と呼ばれる投資信託が存在する．特にテーマ型投資信託は日本の投資家に人気があり，人工知能（AI），ロボティクス，健康など特定の分野を投資先として選んだものとなっている．これらのテーマ型投資信託は，テーマに関連する株式への投資を通じて，テーマの盛り上がりに応じた収益を得ることを目指している．

投資信託会社は，さまざまな種類のファンドを立ち上げ，管理・運用をしている．特に，投資家を引き付けるためには，さまざまな種類のテーマ型ファンドを売り出すことが重要であるとともに，テーマの盛り上がりに対して，適切なタイミングでファンドを売り出していくことが重要である．

テーマ型ファンドの開発には多くの課題がある．テーマ型投資信託を作成するために，投資信託会社は，売れそうな投資信託のテーマを探し，さらに，テーマに

*連絡先：東京大学 大学院工学系研究科 和泉研究室
〒 113-8656 東京都文京区本郷 7-3-1
E-mail: hirano@g.ecc.u-tokyo.ac.jp
HP: https://mhirano.jp/

†本稿は，[Hirano 18, Hirano 19a, Hirano 19b, Hirano 19c] の内容を再構成したものです．

‡特許登録済 (特許第 6596565 号)

§本論文の内容や意見は執筆者個人に属し，いかなる組織の公式見解を示すものではありません．また，特定の商品についての投資の勧誘及び売買の推奨を目的としたものではありません．

関連する銘柄を探さなければならない．加えて，その中から運用において優秀な銘柄を選択してポートフォリオを作成しなければならない．関連する銘柄を探すことは，投資信託を運用するファンドマネージャーにとって，時間的に多大な負担であり，さらに，人手で行うと，本当はテーマに関連のある銘柄を見落としてしまう可能性もある．また，ファンドマネージャーは必ずしもテーマに関して詳しいわけではなく，そういったテーマを取り扱う場合には，事前のリサーチを行うことで対応をしているが，グローバルな投資信託，つまり，海外の銘柄を含むようなファンドの場合，全ての銘柄をカバーすることはほぼ不可能である．そこで，ファンドマネージャーの負担を軽減し，テーマに関連する銘柄を見落としなく抽出するためには，この手順を自動化することが必要である．

近年，様々な技術が発展してきている．特に，テキストマイニングやビッグデータなどの技術は，この分野に応用可能であると考えられる．本論文においては，テキストマイニングに着目して，自然言語処理を用いた，関連銘柄抽出手法を提案する．

2 先行研究

テーマ型投資信託における，ファンドマネージャーをサポートするようなシステムの構築は，現在の日本市場においてはまだまだ先進的な取り組みであり，新規性の高いタスクである．そのため，ここでは，関連する分野における先行研究について述べる．

まず，金融にテキストマイニングを利用している研究をここでいくつか取り上げる．Koppel らは企業の株価のデータを利用して，企業のニュースが良いニュースなのか悪いニュースなのかを分類するテキストマイニングの手法を提案している [Koppel 06]．Low らは “semantic expectation-based knowledge extraction methodology (SEKE)” というニュースなどのテキストから因果関係を抽出する手法を提案しており [Low 01]，単語の概念に関するシソーラスとして，WordNet [Fellbaum 98] を使用している．Schumaker らは金融に関するニュース記事をもとに，株価の予測を行うということを機械学習を利用したアプローチで行なって，その有効性を確かめた [Schumaker 09]．さらに，Ito らは “gradient interpretable neural networks (GINN)” という，金融テキストのセンチメントを可視化するニューラルネットを利用したモデルを提案しており，テキストマイニングにおいて，解釈可能性を追求している [Ito 18b, Ito 18a]．Milea らは，MSCI EURO index の変動予測 (上昇，下降，変化なし) の予測を，European Central Bank (ECB) の発表する文書の文脈から予測する手法を提案している [Milea 10]．Xing らは，アメリカの株式市場における，企業間のつながりを株価の変動の特徴に加えて，テキストマイニングを利用することでより良く抜き出す手法を利用し，アセットアロケーションタスクに組み込むことで，企業間の関係を正確に把握し，ポートフォリオを最適化する手法を提案している [Xing 18]．

次にテキストマイニングを金融の分野に応用している日本語をターゲットとした研究について述べる．Sakai らは日本語のニュースのうち，業績について書かれた文章の中の因果関係を抜き出す手法を提案しており，ブートストラッピング的な手法により，自動的に因果関係を抜き出すための因果を示す手がかり表現を抜き出した [Sakai 07]．Sakaji らはニュース記事から統計的な手法により，自動的に景気動向を示す表現を抜き出す手法を提案した [Sakaji 08]．さらに，Sakaji らは決算短信の中から自動的に稀な因果関係を抜き出してくる手法を提案した [Sakaji 17]．Kitamori らは決算短信の中から業績に関する文章をニューラルネットワークモデルを利用して取り出し，分類する半教師あり学習の手法を提案している [Kitamori 17]．

3 手法

本章では，提案手法について説明する．提案手法はファンドのテーマにおける関連銘柄を選ぶというタスクに対して，ファンドマネージャーをサポートするための手法である．手法概要は図 1 の通りである．まず，テーマの単語をインプットとして入力すると，Word2vec を利用した類似度と企業情報における単語の共起に基づいた類似度を計算し，それらを合わせて最終的な単語の類似度を計算する．単語の類似度を使用して，元のテーマの単語に関連する単語を選び，その単語を含む文を企業情報の中から抽出する．さらに，抽出された文章内に含まれる最終的な関連単語の出現回数とその類似度を加味して企業の類似度を計算する．その結果に基づき，企業の関連度ランキングを作成し，アウトプットとする．この際，抽出の根拠文として，抽出された最終的な関連単語を含む企業情報の文を一緒にアウトプットする．

3.1 Word2vec モデルの構築とそのモデルを利用した類似度計算

まず，テーマの単語の入力を受けた後，Word2vec を利用した類似単語の抽出と類似度計算を行う．Word2vec とは，Mikolov ら [Mikolov 13] によって提案された，単語を多次元の分散表現ベクトルに落とし込む手法である．類似度については，コサイン類似度を使用して計算する．

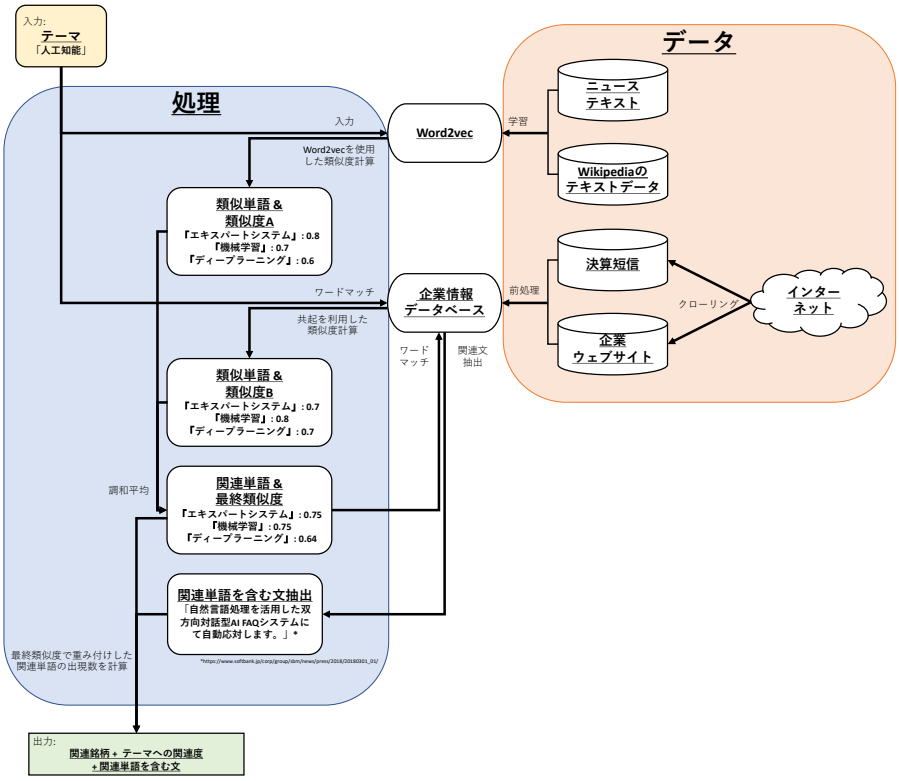

図 1: 手法概要図．例として,「人工知能」を入力とした時を載せている．ただし，数値などは実際のものと異なる．[Hirano 19a] より

Word2vec を利用した類似度計算において，異なるハイパーパラメーターセットで作成した Word2vec モデルを組み合わせて使用している．これは，Word2vec は学習が一意に同じ結果を返すわけではなく，かつ，ハイパーパラメーターによって結果に大きくばらつきが発生するため，それらの影響を緩和するために，異なるハイパーパラメーターセットを利用して学習させたモデルをアンサンブルした．主に，Dimensions，Word window size の異なるパラメーターセットで学習させた．

これらのハイパーパラメーターを使用して学習させたモデルを使用して，類似度を計算した．以下，この類似度を「類似度 A」と呼ぶ．類似度 A の計算方法は以下の通りである．

$$s'_{M_j,word_i} = \begin{cases} c_{M_j,word_i} & (\text{word}_\text{i}\text{が } \text{M}_\text{j}\text{における類似度上位 n 個に入る場合}) \\ 0 & (\text{他}) \end{cases} \quad (1)$$

ここで，$word_i$ は，全ての語彙の中で i 番目の単語，M_i は i 番目の Word2vec モデル，$c_{M_j,word_i}$ はテーマの単語として入力された単語と $word_i$ の M_j のモデルにおけるコサイン類似度，$s'_{M_j,word_i}$ は $word_i$ の M_j における類似度，$s_{A,word_i}$ は $word_i$ の類似度 A である．上記で定義された $s'_{M_j,word_i}$ の調和平均を取ることで，$s_{A,word_i}$ を計算する．さらに，$s'_{M_j,word_i}$ が 0 である場合は $s_{A,word_i}$ も 0 となる．つまり，Word2vec のモデルのいずれかにおいて，$word_i$ が類似度の上位 n 個に入っていない場合に $s_{A,word_i}$ も 0 となる．

この一連のアンサンブル的手法は，Nagata らの研究 [Nagata 18] を参考にしているが，新しく調和平均を取り入れ，拡張を行なった．

3.2 企業情報内における単語共起を利用した類似度計算

本手法において使用している二つの類似度のうち一つは 3.1 節で説明した通り，Word2vec を使用したものであり，もう一つの類似度を本節で説明する．

もう一つの類似度計算は企業情報内における単語共起を利用した類似度計算である．Word2vec は文脈に焦点を当てた手法であり，そのため,「東京」と「大阪」が類似単語として判定されるといった特徴がある．しかし，今回の目的に照らし合わせると，こういった，同じ文脈で使用されるが，関係性の薄い単語が含まれるなどといったことは望ましいとは言えない．そこで，企業情報内における単語共起を利用した類似度計算の手法を新たに提案するとともに，Word2vec と併用するこ

とにした．なお，この手法を利用することで，より下位の概念が強く取れることは，[Hirano 19b] で示している．

企業情報内における単語共起を利用した類似度計算により計算される類似度 B($s_{B,word_i}$) は次の通り計算される．図 2 は計算の具体例を示している．

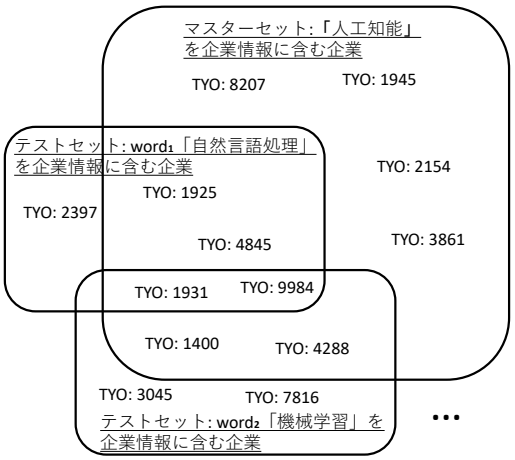

図 2: 類似度 B($s_{B,word_i}$) の計算例. “TYO: xxxx” は東京証券取引所におけるそれぞれの企業の銘柄コードを示している．ここで，$s_{B,word_1} = 4/5 = 0.8$，$s_{B,word_2} = 4/6 = 0.6667$ と計算される．これらの例はあくまで一部であり，実際の結果とは異なる．[Hirano 19a] より．

まず，インターネットを通じてクローリングにより取得した決算短信 (IR) と企業ウェブサイトから取得した企業情報を利用し，テーマの単語を企業情報に含む企業を抽出する．図 2 の例では，テーマの単語「人工知能」に対して，全部で 10 個の企業を抽出できている．(この例はあくまで一部であり，実際の結果とは異なる．) ここで，“TYO: xxxx” は東京証券取引所におけるそれぞれの企業の銘柄コードを示している．以下，このテーマの単語を元に抽出された企業のリストを「マスターセット」と呼ぶ．次に，ほぼ同等の操作を $word_1, word_2, \cdots$ で実施する．違いはターゲットとする単語がテーマの単語ではなく，$word_i$ になることである．つまり，$word_i$ を企業情報に含む企業を抽出していくことになる．図 2 の例では，$word_1$「自然言語処理」を含む企業は 5 つである．(この例もあくまで一部であり，実際の結果とは異なる．) ここで，これらの企業のリストを「テストセット ($word_1$)」と呼ぶことにする．同様にテストセット ($word_i$) を全て計算する．そして，最終的に $s_{B,word_i}$ を次式に従って計算する．

$$s_{B,word_i} = \frac{|\{\text{ マスターセット }\} \cap \{\text{ テストセット } (word_i)\}|}{|\{\text{ テストセット } (word_i)\}|} \quad (2)$$

これは，テストセット ($word_i$) のマスターセットに対する再現率の計算である．例えば，$word_1$「自然言語処理」の場合，5 つの企業がテストセットに含まれ，そのうち，4 つの企業だけがマスターセットにも含まれている．そのため，$s_{B,word_1}$ は $4/5 = 0.8$ と計算される．他も同様に計算されるので，$s_{B,word_2} = 4/6 = 0.6667$ となる．(これらの例もあくまで例であり，実際の結果とは異なる．)

3.3 最終類似度の計算および最終的な関連単語の抽出

以上で説明した二つの類似度を使用して，最終類似度 FS_{word_i} の計算を行う．最終類似度 FS_{word_i} は単なる $s_{A,word_i}$ と $s_{B,word_i}$ の調和平均により計算される．

次に，最終類似度 FS_{word_i} を用いて，最終的な関連単語の抽出を行う．最終類似度 FS_{word_i} の上位 n 個のみを最終的な関連単語とする．以下の説明においては，$word_{f_1}, word_{f_2}, \cdots, word_{f_k}$ を抽出された最終的な関連単語とする．

3.4 関連銘柄の抽出およびその関連度の計算

$word_{f_1}, word_{f_2}, \cdots, word_{f_k}$ に加え，$word_{f_0}$ をテーマの単語とし，$FS_{word_{f_0}}$ を 1 とする．それぞれの企業の企業情報の中から，$word_{f_0}, word_{f_1}, \cdots, word_{f_k}$ それぞれが含まれる文章を抽出し，それぞれの単語の出現回数をカウントする．そして，企業のテーマへの関連度 CS を次式の通り定義する．

$$CS = \sum_{i=0}^{k} FS_{word_{f_i}} \cdot count_{word_{f_i}} \quad (3)$$

ここで，$count_{word_{f_i}}$ は，それぞれの企業の企業情報の中での $word_{f_i}$ の出現回数である．この定義に基づき，各企業のテーマへの関連度が計算され，その関連度に基づいてランキングされる．

4 データと前処理

本章では，本研究において使用したデータと実施した前処理を説明する．本研究においては，日本語の文書を使用している．第一目標が日本市場をターゲットとしたテーマ型投資信託の作成をサポートすることであったからである．日本語の場合，英語などと異なり，形態素間にスペーシングが行われていないため，形態素解析を行う必要がある．

4.1 使用したデータ

本研究において，目的に応じていくつかのデータを使用した．目的は主に二つで，Word2vec の学習用および企業情報としてである．Word2vec 学習用として，ライブドアニュースコーパス[1]，Wikipedia 日本語記事 (version 21-Jun-2018 22:09) [2]，日本経済新聞 (1990–2015 and 2017; 2016 年分は技術的問題から使用していない) を利用した．これらのデータを元に，1,147,973 語彙，1,809,736,365 文字を含むテキストデータを Word2vec の学習に利用した．

企業情報としては，2012/12/9 — 2018/5/11 の決算短信，全 90,813 ファイル (PDF) を日本取引所グループ適時開示情報閲覧サービス (TDnet) [3] より取得した．企業ウェブサイトのデータとしては，2018/6/6 — 2018/6/25 の期間に，2,293,460 ファイルのみ (703,699 PDF，1,472,317 HTML，その他) を取得した．

これらの企業情報はインターネットを通じて取得した．本研究においては，これらの多種のデータを組み合わせることで，関連銘柄を抽出している．

4.2 実施した前処理

日本語の文書においては形態素解析が必要である，日本語の形態素解析器として，KyTea [Neubig 11] や JUMAN++ [Morita 15] などがあげられる．本研究においては，もう一種類の MeCab (version 0.996)[4]を使用した，また，MeCab 用の追加の日本語辞書として，NEologd[5]を使用した．

MeCab および NEologd を使用して，全てのテキストデータを形態素に分解してから使用している．Word2vec の学習においては，形態素解析して，形態素間にスペースを入れたデータを Word2vec の学習用テキストとして使用した．一方，企業情報は検索を早くするために，形態素解析後のテキストを，どの企業の情報なのかやデータのソースなどとともにリレーショナルなデータベースに保存して使用した．特に，収集した企業情報は 600GB 程度あり，実験において，リレーショナルな形で保存されていないと照会が非常に困難であるため，リレーショナルなデータベースでの保存を採用した．

5 実験と結果

本章においては，本手法のハイパーパラメーターチューニングと評価実験およびそれらの結果について述べる．

各節ごとでの説明の前に，実験で共通で使用している評価用データについて述べる．評価用データとは，提案手法等を使用して出てきた結果を評価するために使ったデータである．評価用に使用したテーマは「美容」,「育児」,「ロボット」,「娯楽」の 4 つである．経験のあるファンドマネージャー 4 名のタグ付けデータを使用し，それを元に評価用データを作成した．TOPIX500 の中からランダムに選ばれた 100 銘柄を使用し，タグ付けを行なってもらった．タグ付けの基準は同様で，以下の通りである．

0. 全く関係ない
1. あえて言えば関連している
2. この企業の事業の一部に関係がある
3. 非常に関係があり，この企業の代表的な事業だ

ただし，一つだけ異なるのは，ファンドマネージャーは自分のタグ付けに自信がない場合には，その旨のタグ付けを行うことができる．

表 1: ファンドマネージャーのタグ付けの例．“+” は銘柄のタグ付けに自信がないことを示す．

銘柄コード	FM1 タグ	FM1 自信なし	FM2 分類	FM2 自信なし
4544	1	+	0	
4555	1	+	0	
4578	1	+	0	
4661	2		0	
4665	1		3	
4676	1		0	

表 1 はファンドマネージャーによるタグ付けの例である．例えば，銘柄コード 4544 のみらかホールディングスに対して，ファンドマネージャー 1 は自信がないものの,「あえて言えば関連している」とタグをつけた．一方で，ファンドマネージャー 2 は「全く関係ない」とタグをつけた．このファンドマネージャー 4 名によるタグ付けに基づいて，評価用データを作成する．関連銘柄の判定は非常に難しいタスクである．(詳細については [Hirano 19c] を参照．) そのため，評価用データにおいては，1 名以上のファンドマネージャーが関連のある銘柄だと判断した場合には，関連銘柄とすることとした．これは，人間が行う以上，知識不足や見落としにより，関連のない銘柄だと判定してしまう可能性が充分に存在するからである．

そこで，ファンドマネージャーによるタグ付けを評価用データに変換する際の最終的な基準を次のように定めた．

[1] https://www.rondhuit.com/download.html#ldcc
[2] https://dumps.wikimedia.org/jawiki/latest/
[3] https://www.jpx.co.jp/equities/listing/tdnet/
[4] http://taku910.github.io/mecab/
[5] https://github.com/neologd/mecab-ipadic-neologd

表 2: ファンドマネージャーによるタグ付けの評価用データへの変換例．“FM” はファンドマネージャーを，“+” は自信がないことを示す．

銘柄	FM によるタグ付け				評価用データ	
	FM1	**FM2**	**FM3**	**FM4**	関連銘柄	関連度レーティング
A	1	3	2	1	✓	$(1+3+2+1)/4 = 1.7500$
B	0	1	0	0	✓	$(0+1+0+0)/4 = 0.2500$
C	0	1(+)	0	0	-	$(0+1\times0.5+0+0)/3.5 = 0.1429$
D	2	1(+)	0(+)	0	✓	$(2+1\times0.5+0\times0.5+0)/3 = 0.8333$
E	0	0	1(+)	0(+)	-	$(0+0+1\times0.5+0\times0.5)/3 = 0.1667$
F	0	0	1(+)	1(+)	-	$(0+0+1\times0.5+1\times0.5)/3 = 0.3333$

- 自信のないファンドマネージャーを除き，1 名以上のファンドマネージャーが関連銘柄だと判定した (1,2,3 のいずれかにタグ付けした) 場合は関連銘柄としてみなす．
- 4 名のファンドマネージャーの 0-4 の関連度のタグ付けを算術平均したものをその銘柄の関連度レーティングとする．ただし，自信のないファンドマネージャーが存在する場合は，平均を取る際にその人の重み付けを 1 ではなく，0.5 とする．

表 2 はファンドマネージャーによるタグ付けから評価用データへの変換例である．銘柄 A と B は典型的な例である．全てのファンドマネージャーが銘柄 A は関連銘柄とみなしている．そのため，評価用データにおいても，関連銘柄としている．銘柄 B も評価用データにおいては関連銘柄となっている．これは，1 名のファンドマネージャーが関連銘柄だとみなしているからであり，すでに述べたとおり，1 名ファンドマネージャーのみが関連銘柄とみなしている場合でも，他のファンドマネージャーが知識不足や見落としなどから関連銘柄としていない可能性が充分にあると考えられるからである．一方で，銘柄 C は少し異なる．1 名のファンドマネージャーが関連銘柄としているが，そのファンドマネージャーは自信がないとしている．このような場合には，評価用データにおいては，関連銘柄と判定しない．さらに，銘柄 D-F はもっと特殊なケースである．銘柄 C においても同様ではあるが，自信のないファンドマネージャーは 0.5 名分としてカウントして平均をとったものを関連度レーティングとしている．さらに，銘柄 E と F においては，関連銘柄とタグをつけているファンドマネージャーが全員自信がない場合のため，評価用データにおいては関連銘柄とは判定していない．

このように基準を定めたものの，自信がないとタグをつけられたデータは $28/1200 = 2.3\%$ しかなく，実際には 2 名以上が自信がないことを示すタグを同一テーマ，同一銘柄につけた銘柄 D-F のようなケースは存在しなかった．

5.1 ハイパーパラメーターチューニング

より良いハイパーパラメーターを探すために，ハイパーパラメーターチューニングを行なった．企業情報のデータソース，3.1 節で説明した複数の Word2vec のモデルをアンサンブルするフェーズにおける上位何個を取るかという値 (Hyperparameter1 という)，3.3 節で説明した，最終類似度の計算における上位何個を取るかという値 (Hyperparameter2 という) を変更して，実験をし，その結果からチューニングを行なった．ハイパーパラメーターのチューニングの方式はグリッドサーチであり，それぞれのパラメーターに対していくつかの離散的な値を使用して，その全ての組み合わせを実験し，その中で良い結果を出したパラメーターセットを使用するという手法である．グリッドサーチで使用したそれぞれのパラメータに対するグリッドは以下の通りである．

- 企業情報のデータソース: (1) 決算短信のみを使用，(2) 企業ウェブサイトからのデータのみを使用，(3) 決算短信および企業ウェブサイトからのデータの両方を使用
- Hyperparameter1: top-10, top-20, top-50, top-100, top-200, top-500, top-1000, top-2000
- Hyperparameter2: top-5, top-10, top-20, top-50, top-100, top-200, top-500, top-1000

評価用データとしては,「美容」,「育児」,「ロボット」,「娯楽」があるが，そのうち,「美容」,「育児」,「娯楽」の三つをハイパーパラメーターのチューニングに利用し，残りの一つの「ロボット」を性能評価用に使用した.「ロボット」を性能評価用に使用した理由は，事前の実験で，最も低い結果を出していたからである．

ハイパーパラメーターのチューニングにあたっては，結果の良し悪しをいちいち人間が判断することは不可能であり，何かしらの「良さ」の指標を利用しなければならない．そこで，本実験においては，F1 による

チューニングを採用することとした．結果として，採用したハイパーパラメーターセットは企業情報のデータソースが決算短信のみ，hyperparameter1 が top-500，hyperparameter2 が top-50 となった．

5.2 テストデータでの性能評価

前節で定めたハイパーパラメーターを採用し，残っている一つの「ロボット」の評価用データを使用して，性能評価を行なった．比較として，企業情報として決算短信のみを使用し,「ロボット」という単語をそれぞれの企業情報に単純照合をすることで企業を抽出するという実験を行なった．結果は表 3，4，5 の通りである．

表 3: 「ロボット」の評価用データとチューニング後のハイパーパラメーターを使用した場合の提案手法の結果に対する混同行列

		評価用データにおいて 関連銘柄	非関連銘柄
実験結果	関連銘柄	48	34
	非関連銘柄	9	9

表 4: 企業情報として決算短信のみを使用して,「ロボット」という単語をそれぞれの企業情報に単純照合をすることで企業を抽出した場合の結果に対する混同行列 (比較用)

		評価用データにおいて 関連銘柄	非関連銘柄
実験結果	関連銘柄	9	5
	非関連銘柄	48	38

表 5: テスト用テーマ「ロボット」における，提案手法と比較手法の Precision，Recall，F1，Accuracy．比較手法は企業情報として決算短信のみを使用して,「ロボット」という単語をそれぞれの企業情報に単純照合をすることで企業を抽出するという手法である．

	Precision	**Recall**	**F1**	**Acc.**
提案手法	0.5854	0.8421	0.6906	0.5700
比較手法	0.6429	0.1579	0.2535	0.4700

表 5 を見れば明らかであるが，表 3 と 4 を比べると，比較手法に比べて，より多くの銘柄を関連銘柄として抽出しているため，Precision が低下している．一方で，圧倒的に Recall が向上しており，その結果，F1 も向上している．また，単純な結果の正確さを測る Accuracy も向上しており，提案手法が良い結果を出していることがわかる．

6 考察

本研究は，テーマ関連銘柄の抽出を支援するシステムの構築を目標にしたものであった．[Hirano 19c] で示しているが，これは，経験のあるファンドマネージャーにとっても難しいほどのタスクであった．そのため，そもそも評価用データがどれだけ正しいのか，という点については多少の疑問が残る．しかし，表 5 を見る限りでは，基本的な手法を利用した比較手法よりは良い結果を出しており，提案手法の一定の有効性が示されたと考えられる．

今回の結果は企業ウェブサイトからのデータはなくても，決算短信だけでも充分であるという結果になっている．しかしながら，企業ウェブサイトのデータを使用した場合には，一つの単語で抽出できる企業数が多いこともわかっている．これは，決算短信だけでなく，企業ウェブサイトからのデータも使った場合には，少ない数の関連単語のみを使用することで，充分な関連銘柄を抽出していた一方で，決算短信だけを企業情報で使用する場合には，より多くの関連単語を採用することで，データの少なさをカバーすることができているのかもしれない．しかし，実験で実際に使用した 4 つの単語はすでに広く普及している，一般的な単語であった．そういった一般的単語は決算短信には頻出する一方で，比較的新しい単語は決算短信には出現しにくい可能性がある．そのため，今後の課題として，提案手法において，企業ウェブサイトからのデータを使用しなかった場合に，比較的新しい単語においても充分な精度を発揮するのかを確認しなければならない．

今回の実験における，ハイパーパラメーターのチューニングはある種の教師あり学習のような形のグリッドサーチで行なった．教師データとしては，ファンドマネージャーにタグ付けしてもらったデータを元に作成した評価用データを利用したが，このデータがどれだけ正しいかわからないだけでなく，作成も人手で行うため，時間のかかる作業であり，多くの評価用データを作成するのは困難であった．そのため，評価に使用したデータのテーマは限られており，これでは充分な結果とは言えない可能性もある．もちろん，より正確で，多くの評価用データを作成することができれば問題はないが，限界がある．さらに，現在は日本株だけを対象にしていたが，これを世界中の株を対象にすると，評価用データの作成の難易度はもっと上がると考えられる．そのため，教師なし学習または半教師あり学習のような手法を作成しなければ，どんなテーマに

も対応できるような手法とはならないと考えられる．

今後の展望としてだが，現時点では，単純にテーマへの関連度のみを考慮してランキングを作成していたが，実際にはもう少し他の軸を加えてランキング付けしたいという需要もある．例えば，テーマにおける代表性や企業の大きさを加味したランキングが欲しいという意見もファンドマネージャーから聞いた．そのため，様々な指標を取り入れられるようなスキームを作成することも必要かもしれない．また，銘柄をテーマ型投資信託に採用し，運用を行う際には，ポートフォリオ運用をすることになるので，同じような株価変動をする銘柄ばかりを組み込むことは望ましくない．それは，単に価格変動だけではなく，似たような事業を行なっている企業を同時に取り込みことは，結果として，同じような価格変動を引き起こす可能性を秘めている．そのため，価格変動や企業情報を主成分分析したり，ベクトル化することで，様々な需要にうまく適用できるようにする技術の開発などが必要であると考える．

謝辞

本研究は，大和証券グループと東京大学大学院工学系研究科により開設された社会連携講座「次世代運用テクノロジー」における，大和証券投資信託委託株式会社および株式会社大和総研との共同研究の一部であり，大和証券グループの支援を受け，行われました．大和証券グループ各社の多くの方にご支援をいただきました．この場を借りて御礼申し上げます．

参考文献

[Fellbaum 98] Fellbaum, C.: *WordNet: An Electronic Lexical Database*, The MIT Press (1998)

[Hirano 18] Hirano, M., Sakaji, H., Kimura, S., Izumi, K., Matsushima, H., Nagao, S., and Kato, A.: Selection of Related Stocks using Financial Text Mining, in *Proceedings of 18th IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 191–198, Singapore, Singapore (2018)

[Hirano 19a] Hirano, M.: *Extraction of Related Stocks for Themed Mutual Funds using Text Mining*, Graduation thesis, The University of Tokoy (2019)

[Hirano 19b] Hirano, M., Sakaji, H., Kimura, S., Izumi, K., Matsushima, H., Nagao, S., and Kato, A.: 文書内における単語の共起を利用した上位下位概念の推定, 言語処理学会第 25 回年次大会, pp. 597–600, Nagoya, Aichi, Japan (2019), The Association for Natural Language Processing

[Hirano 19c] Hirano, M., Sakaji, H., Kimura, S., Izumi, K., Matsushima, H., Nagao, S., Kato, A., Hirano, M., Sakaji, H., Kimura, S., Izumi, K., Matsushima, H., Nagao, S., and Kato, A.: Related Stocks Selection with Data Collaboration Using Text Mining, *Information*, Vol. 10, No. 3 (2019)

[Ito 18a] Ito, T., Sakaji, H., Izumi, K., Tsubouchi, K., and Yamashita, T.: GINN: gradient interpretable neural networks for visualizing financial texts, *International Journal of Data Science and Analytics*, pp. 1–15 (2018)

[Ito 18b] Ito, T., Sakaji, H., Tsubouchi, K., Izumi, K., and Yamashita, T.: Text-Visualizing Neural Network Model : Understanding Online Financial, in *Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2018)*, pp. 247–259, Melbourne, Australia (2018)

[Kitamori 17] Kitamori, S., Sakai, H., and Sakaji, H.: Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning, in *Proceedings of 2017 IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2017)*, pp. 67–73, Honolulu, Hawaii, USA (2017)

[Koppel 06] Koppel, M. and Shtrimberg, I.: Good News or Bad News? Let the Market Decide, in *Computing Attitude and Affect in Text: Theory and Applications*, pp. 297–301, Springer Netherlands (2006)

[Low 01] Low, B.-T., Chan, K., Choi, L.-L., Chin, M.-Y., and Lay, S.-L.: Semantic expectation-based causation knowledge extraction: A study on Hong Kong stock movement analysis, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2001)*, pp. 114–123 (2001)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, in *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pp. 1–12, Scottsdale, Arizona, USA (2013)

[Milea 10] Milea, V., Sharef, N. M., Almeida, R. J., Kaymak, U., and Frasincar, F.: Prediction of the MSCI EURO index based on fuzzy grammar fragments extracted from European Central Bank statements, in *2010 International Conference of Soft Computing and Pattern Recognition*, pp. 231–236 (2010)

[Morita 15] Morita, H., Kawahara, D., and Kurohashi, S.: Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 2292–2297, Lisbon, Portugal (2015)

[Nagata 18] Nagata, R., Nishite, S., and Ototake, H.: A Method for Detecting Overgeneralized Be-Verb based on Subject-compliment Identification [published in Japanese], *The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018 (JSAI 2018)* (2018)

[Neubig 11] Neubig, G., Nakata, Y., and Mori, S.: Pointwise Prediction for Robust , Adaptable Japanese Morphological Analysis, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pp. 529–533, Portland, Oregon, USA (2011)

[Sakai 07] Sakai, H. and Masuyama, S.: Extraction of Cause Information from Newspaper Articles Concerning Business Performance, in *Proceedings of the 4th IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI 2007)*, pp. 205–212 (2007)

[Sakaji 08] Sakaji, H., Sakai, H., and Masuyama, S.: Automatic Extraction of Basis Expressions That Indicate Economic Trends, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)*, pp. 977–984 (2008)

[Sakaji 17] Sakaji, H., Murono, R., Sakai, H., Bennett, J., and Izumi, K.: Discovery of Rare Causal Knowledge from Financial Statement Summaries, in *The 2017 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr 2017)*, pp. 602–608 (2017)

[Schumaker 09] Schumaker, R. P. and Chen, H.: Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System, *ACM Transactions on Information Systems*, Vol. 27, No. 2, pp. 12:1–12:19 (2009)

[Xing 18] Xing, F., Cambria, E., and Welsch, R. E.: Growing Semantic Vines for Robust Asset Allocation (2018), https://ssrn.com/abstract=3275132