決算報告書数字表現コーパスの設計と構築

村上 浩司
楽天技術研究所 楽天株式会社
koji.murakami@rakuten.com

1 はじめに

近年，経済や金融に関する研究が数多く行われており，様々なデータや文書が用いられている．株価予測の研究においては過去の株価データとともに，Twitter[15, 11]，ニュース記事 [13, 2]，収支報告 (Earning Call)[12]，8-K（米国における財務に関する臨時報告書）[4]，10-Q（四半期報告書）や 10-K（年次報告書）[1, 5]，ヤフーファイナンス掲示板 [7] などが言語資源として利用されている．また他に，株価からの市況コメント生成 [6] ではニュース記事，金融リスクマネジメント [3] においては企業の開示情報，辞書構築 [16] では，FOMC（Federal Open Market Committee:米国連邦公開市場委員会）[1]議事録も利用されている．Xing らは金融情報を対象とした言語処理やテキストマイニング利用される言語資源を (1) 企業が開示する情報，(2) 研究機関が作成する金融レポート，(3)Wall Street Journal(WSJ) や Financial Times のような専門刊行物，(4)Bloomberg や Thomson Reuter などのニュースポータル，(5) 掲示板，(6) ソーシャルメディア，の 6 種類に分類した [14]．表 1 にそれぞれの言語資源の特徴を示す．

株価予測研究の多くは，ある時点での対象企業に関するニュース記事などから極性を判定し，過去の株価変動のパタンと照合して予測結果を出力するものが多い．この場合，主に着目しているのはセンチメントと呼ばれる市場心理であり，企業の財務状況などを示すファンダメンタルズを直接は考慮していない．より精度の高い企業の株価予測のためには，企業の決算，財政状況や企業活動の情報を適切に整理，利用する必要がある．こうした情報は個々の企業から提供される決算短信や有価証券報告書などの決算報告書に書かれている，貸借対照表や損益計算書をはじめとした財務諸表や，企業の状況などの説明文から読み解くことができる．

表 1: 金融に関する言語資源とその特徴

種類	文書長	論調	発行頻度
企業の開示情報	長	主観的	低
金融レポート	長	客観的	低
専門刊行物	短-長	客観的	中
ニュースポータル	中	主/客観的	低-高
掲示板	短	客観的	高
ソーシャルメディア	短	主観的	高

[1]https://www.federalreserve.gov

先に挙げた決算報告書を利用した株価予測の研究でも，ニュース記事を利用した場合と同様，基本的に考慮する情報は単語の極性やキーワードの有無とその出現コンテキストなどである．しかしながらこうした言語情報だけではなく，決算報告書には多くの重要な数字表現も記載されている．財務諸表などの表形式で明示的に書かれた収入や利益の数字も重要ではあるが，例えば「連結子会社数の変動」，金融機関であれば「ATM 設置数の増減」といった財務諸表中の数字を直接説明しない文中の数字表現も，その企業の状況の位置付けのためには重要な情報になりうる．

そこで我々は，企業の情報を詳細に整理することを目的として，上場企業が提供する決算報告書のうち有価証券報告書を対象とし，自然文中に出現する数字表現の勘定科目分類を行い，それぞれのクラスのタグ付与することでコーパスの構築を行う．また固有表現の種類としての数値表現について考察し，固有表現抽出のサブタスクとしての可能性を検討する．

2 有価証券報告書内の数字表現の構造化

有価証券報告書とは，金融商品取引法で定められる事業年度ごとに各企業が作成する企業内容の外部への開示資料であり，EDINET(Electronic Disclosure for Investors' NETwork)[2]への電子提出が義務付けられている．提出される資料のうち財務諸表本表については，XML を拡張した XBRL(eXtensible Business Reporting Language) 形式で作成される．これにより表の内部の情報に関しては構造化されるが，表内の数字について説明している文もしくは事業の状況などを述べた文中に現れる数字表現に対して構造化は行われない．

2.1 先行研究との比較

有価証券報告書に出現する数字表現には，大きく分けて個数や割合などを示すものと，時間を表すもの，モノ

[2]http://disclosure.edinet-fsa.go.jp/

表 2: アノテーション対象の上場企業とその業種

業種	企業数	作業数	作業比率	企業名
サービス業	439	3	0.68%	楽天，リクルートホールディングス，日本郵政
不動産業	133	1	0.75%	住友不動産
商業	673	5	0.74%	三菱商事，伊藤忠商事，丸紅，イオン*，ファーストリテイング
建設業	163	1	0.61%	大和ハウス*
水産・農林業	11	1	9.09%	サカタのタネ
製造業	1,469	14	0.95%	武田薬品工業，出光興産，ソニー，富士フィルムホールディングス トヨタ自動車，キヤノン*，ブリヂストン*，JXTG ホールディング 任天堂，日立製作所，東洋紡，新日鐵住金，デンソー*，資生堂
運輸・情報通信業	546	4	0.73%	ソフトバンクグループ，ヤフー，日本電信電話，東日本旅客鉄道*
金融・保険業	177	3	1.69%	三菱 UFJ ファインシャル・グループ，かんぽ生命保険，セブン銀行
鉱業	6	1	16.57%	日鉄鉱業
電気・ガス業	24	1	4.17%	関西電力*
合計	3,641	34	0.94%	

の名前の一部としての表現が存在する．様々な数字表現に対してその概念のクラスを定義している分類体系に関根の拡張固有表現階層がある [10]．この階層中の金額表現，割合表現，個数に対応する数字表現が有価証券報告書には多く存在する．本研究の目的は数字表現の勘定科目分類であり，例えば科目の異なる営業利益と資金流出を同じ金額表現クラスにまとめることはできない．しかしながら科目の概念的なクラス分類もまた将来的に必要と考えられることから，科目の他に拡張固有表現階層中の数値表現 (Numex) のタグも同時に付与する．

時間を表す数字表現に対する分類体系には拡張固有表現階層の他に時間情報タグ付け基準 TimeML[9] があり，これを使った TimeBank コーパスが整備されている [8]．決算報告の場合，例えば日付などは明示的に記載され，それほど複雑な事例がないと予想されることから，時間表現のタグ付けは拡張固有表現階層でシンプルに定義される時間表現（Time_Top）体系に従うこととする．

モノの名前を表す数字が含まれる表現には方式名や「S&P500」などの製品名が考えられるが，これらは適宜拡張固有表現階層中で最も相応しいタグを付与する．

2.2 業種と対象企業

東京証券取引所（第一部，第二部）及び新興企業市場（ジャスダック，マザーズ），Tokyo Pro Market で取引される企業の合計数は 2018 年 11 月 30 日現在で合計 3,641 である．[3] 上場企業は多種多様の業種から構成されている．これらを分類する目的でいくつかの分類基準が用いられている．幅広く利用されている体系には証券コード協議会が定める中分類と大分類がある[4]．中分類は，東証 33 業種とよばれる 33 業種であり，これらを再編した 17 業種（TOPIX-17）も幅広く利用されている．大分類は中分類を 10 業種へのまとめたものである．またその他に，日本標準産業分類による 2 業種，36 業種，256 業種への分類も存在する．

本論文ではこれらの分類基準のうち，大分類に準拠し企業を業種に分類する．我々はアノテーション作業を行う企業の選別基準として，各業種から少なくとも 1 社，またそれぞれの業種の企業の 1% を目安とした．業種内の企業の選定は，できるだけ時価総額が大きい会社であること，主要業務が重複する企業を選ばないようにした．また，アノテーション作業の時間差の影響から，会計年度が若干異なる．その結果，34 社が作業対象として選ばれた．表 2 に作業対象の企業と業種について示す．基本的に 2017 年に提出された報告書が対象であるが，“*” が付いた企業は 2018 年に提出された報告書が対象である．

3 コーパス作成

3.1 文書収集とデータ準備

作業対象の企業の有価証券報告書は，EDINET 経由でアーカイブを取得した．フォーマットは PDF と XBRL，HTML の 3 形式で作成されている．有価証券報告書の内容は企業の種類によって章立てが定義されており，財務諸表等は「経理の状況」に記載される．報告書には財務諸表の他にも，数多くの表が記載されている．こうした表は本研究の目的である企業の情報の整理のためには不可欠であるが，そのための方策が自然文からの情報抽出とは異なることから，ここでは対象にはしない．我々は，最も自然文による記述量が多い「事業の状況」の章に着目し，アノテーション作業を行う．

我々は以下の手順で前処理を行い，作業用のデータを準備する．まず (i) 各企業の有価証券報告書の「事業の状況」が書かれた HTML からタグを除去しすべての文を取得後，半角数字を全角に変換，(ii) ローマ数字，漢数字，桁間のカンマからなる連続した数字列を抽出，(iii) 原文，数字列とともに，前後 3 単語ずつをコンテキストとして保存することで作業データを作成する．この前処

[3] http://www.jpx.co.jp/markets/statistics-equities/misc/tvdivq0000001vg2-att/data_j.xls

[4] https://www.jpx.co.jp/sicc/category/ct_chart.html

表 3: 業種ごとの数字表現，及び拡張固有表現クラスの分布

業種	サービス	不動産	商業	建設	水農林	製造	通信	金融	鉱業	電気	合計
数字表現数	1,585	317	1,575	245	191	5,083	2,018	1,818	176	219	13,227
[**ENE**]											
末端ノード数	22	12	30	14	12	43	32	27	6	11	57
時間	213	80	195	21	12	1,034	535	211	4	57	2,362
金額表現	379	108	700	83	87	1,491	462	372	100	66	3,848
割合表現	206	33	246	28	22	593	234	303	24	38	1,727
個数	185	23	61	8	8	219	134	77	1	5	721
上位 5 種	1,420	278	1,451	213	177	4,430	1,755	1,481	175	206	11,526
上位 5 種 (%)	89.6%	87.7%	92.1%	86.9%	92.7%	87.2%	87.0%	81.5%	99.4%	94.1%	87.1%

理により，34 企業の報告書から 13,227 箇所の数字表現が得られた．ここで抽出された数字列は数量を表す場合でも単位を含まず，また製品名の一部であることも考えられる．最終的には単位を含んだ表現や製品名全体に対してクラスのタグ付けを行うが，科目を特定する作業の段階では数字列とコンテキストから判断する．

3.2 分類タグの設計

一般的にコーパスの構築手順は，(i) 抽出したい情報を定義し分類（クラスのリスト作成），(ii) 作業対象の文書群を準備，(iii) 文書に対して定義した情報が存在すればクラスのタグ付与（アノテーション作業），となる．しかしながら本研究の場合の作業対象は，金融庁により定義された既知の勘定科目の数字表現のみではなく，企業の業種や個々の企業の事業内容に依存する未知の数字表現も考慮する必要がある．そのため，できるだけ幅広い種類の企業の報告書に対しての作業が重要である．

文中の勘定科目に関する数字表現の場合は，金融庁が定義する EDINET タクソノミ要素リスト[5]と勘定科目リスト[6]に準拠する形でラベルを付与する．勘定科目にない数字表現の場合は，34 企業の分析作業後にクラスをまとめることとする．

3.3 アノテーション

対象の 13,227 箇所の数字表現への勘定科目及び拡張固有表現階層に基づくクラス名付与，勘定科目以外の数字表現へのクラスの分析を行った．表 3 に結果を示す．

拡張固有表現に関して，実際のクラスのタグ付け作業は末端クラスの名称を利用したが，表中のクラスについては定義の中の第一階層の名称にまとめた．すべての業種に出現したクラスは時間，金額表現，割合表現，個数と序数であった．各業種において頻度の高い上位 5 つのクラスの数字表現の合計は，全体の 9 割程度であることがわかる．金融においてのカバレッジが少し低いが，これは他の業種に出現しない金融分野に特有な用語が別のクラスとして中頻度で出現していたのが要因であった．

表 4: 勘定科目にない数字表現の大まかな分類

代表表現	#	例数	主な表現
割合など	34	529	出資比率，為替換算レート
利益など	23	1,058	営業外損益，ネット有利子負債
残高など	17	239	年金資産，拠出額，連結売上高
～数	13	487	会社数，株式数，創業年数
年月日	6	2,400	年度，〆日，期間始，期間終
その他	51	4,659	住所，生産量，格付け，利回り

勘定科目の付与は，リストを照合しつつ進めるが必ずしも文中の表現が一致するとは限らないため，文中の表現に対しクラス候補を記述することとする．例えば「不動産・ホテル事業の総資産は…」の場合，クラス候補として「総資産_不動産・ホテル事業」，また「長期での期待収益率は」では「期待収益率_長期」のような，金融関連用語とその主体・実体などをペアとした形で記述する．数字表現全体に対してクラス候補を付与したところ，全体で 1,786 種類になった．このうち 721 種類はペアのどちらかが勘定科目リスト中の 157 科目と一致した．一致した勘定科目の数字表現数は 3,149 個であった．次に一致しなかった 1,064 のクラス候補のうち，金融関連用語のみを抽出してまとめ上げた．その結果，勘定科目以外のクラス分けが必要な用語は 469 種類であった．このうち頻度が低いものは特定の業種に特化した表現であると思われる．勘定科目でない数字表現の総数の 90%以上は頻度 6 以上の金融関連用語でカバーされ，このときの用語は 144 であった．これらを大まかに分類した結果を表 4 に示す．

勘定科目に一致しない数字表現のうち，おおよそ 25%(2,400/9,372) が年月日，その他に属する数字表現はほぼ 50%(4,659/9,372) であることから，財務分析等に用いられる金融・経済用語以外にも多くの数字表現が存在することがわかる．

3.4 検討中の課題

対象とする「事業の内容」の章において，収入や支出，利益などは数字表現単独ではなく，次の例のように詳細な説明のために会計年度の情報や割合を伴って記述されることが多い．

[5]https://www.fsa.go.jp/search/20180228/1e_ElementList.xls
[6]https://www.fsa.go.jp/search/20180228/1f_AccountList.xls

(1) 投資活動による収入は，前連結会計年度に比べ１６４８億円増加（同９％増）し，１９８７０億円となりました．

(2) 長期期待収益率が０．５％低下した場合，翌連結会計年度の費用は約４９億円増加します．

(3) 当連結会計年度におきましては，営業損失が１００５百万円増加しております．

例 (1) では，当連結会計年度において 19,870 億円が収入であること，さらに前連結会計年度との差分が金額では 1,648 億円の増加，比率では 9%増加であるという 3 種類の情報が記載されている．報告書中には当年度のことだけでなく，翌年度の会計年度に対する予測も例 (2) のように書かれることがある．そのため，付与するタグは例 (1) では “<revenue period=”0” ydiff=”-1” type=”yen” flag=”plus”>１６４８億円</revenue>” や “<revenue period=”0” ydiff=”-1” type=”ratio” flag=”plus”>９．０％</revenue>”，例 (2) では “<cost period=”+1” ydiff=”-1” type=”yen” flag=”plus”>４９億円</cost>” のように記述することで会計年度などを考慮することができる．例 (3) では営業損失について述べているが，勘定科目では「営業利益又は営業損失（△）」と定義されていることから，“<operating_profit period=”0” diff=”-1” flag=”minus”>１００５百万円</operating_profit>” とすることで，額や比率の増減だけでなく利益か損失かを表現することができる．

これまで作業を行なった報告書中の勘定科目のカバレッジをそれぞれの業種で調査した．金融庁の提供する勘定科目リストは，本研究で利用した事業に基づいた分類である 10 業種を利用しておらず，共通の勘定科目多くなるような業種で分類された 23 種類を用いている．これらを手作業で大分類に近似するよう 6 種類に分類した結果を表 5 に示す．勘定科目リストで定義されている科目数は全体で 3,046 種類である．34 社への作業でリストとマッチしたのは 157 種類であり，そのうち 95 種が業種頻度が 1 ではないものであった．科目は必須もしくは選択と定義されていないため単純には正しいカバレッジは計算できないが，単純計算で約 20%(=95/(3,046-2,570)) と計算できる．これにより，表 2 の作業企業数を向上させる必要があることがわかる．

表 5: 勘定科目の大分類へのマップと科目数

業種	数	科目数 (業種頻度=1)
製造業	2	1,550 (1,101)
建設業	2	329 (91)
運輸・情報通信業	3	481 (237)
電力・ガス業	2	376 (204)
金融業	12	1,234 (845)
サービス業	2	223 (92)
合計	23	3,046 (2,570)

4 まとめと今後の課題

本稿では，有価証券報告書内の自然文中の数字表現に対して勘定科目及び拡張固有表現のクラスを付与するコーパスの設計と構築について報告した．現在 34 企業の報告書に対してクラス付与のための分析を進めているが，捉えられるクラスの数がまだ少ないためさらに多くの企業の報告書を作業対象にすることが次の大きな課題である．また，実際の文書には参照する勘定科目の表記揺れもしくは類義表現が数多く確認されたため，類義語辞書の構築を進める予定である．

参考文献

[1] Matthew Butler and Vlado Keselj. Financial forecasting using character n-gram analysis and readability scores of annual reports. *Advances in artificial intelligence*, pp. 39–51, 2009.

[2] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Knowledge-driven event embedding for stock prediction. In *In Proc. of the 26th International Conference on Computational Linguistics (COLING 2016)*, pp. 2133–2142, 2016.

[3] Sven S. Groth and Jan Muntermann. An intraday market risk management approach based on textual analysis. *Decision Support Systems*, Vol. 50, No. 4, pp. 680–691, 2011.

[4] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In *In Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[5] Feng Li. The information content of forward-looking statements in corporate filings - a naive bayesian machine learning approach. *Journal of Accounting Research*, Vol. 48, pp. 1049–1102, 2010.

[6] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. Learning to generate market comments from stock price. In *In Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1374–1384, 2017.

[7] Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *In Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics*, pp. 1354–1364, 2015.

[8] James Pustejovsky, Jose Castano, Robert Ingria, Reser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. The timebank corpus. In *In Proc. of Corpus Linguistics 2003*, pp. 647–656, 2003.

[9] James Pustejovsky, Jose Castano, Robert Ingria, Reser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. Timeml: Robust specification of event and temporal expression in text. In *In Proc. of IWCS-5, Fifth International Workshop on Computational Semantics*, 2003.

[10] Satoshi Sekine. Extended named entity ontology with attribute information. In *In Proc. of the 5th International Conference on Language Resources and Evaluation*, pp. 52–57, 2008.

[11] Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. Exploiting social relations and sentiment for stock prediction. In *In Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 1139–1145, 2014.

[12] William Yang Wang and Zhenhao Hua. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *In Proc. of the 52th Annual Meeting of the Association for Computational Linguistics*, pp. 1155–1165, 2014.

[13] Boyi Xie, Rebecca J. Passonneau, Leon Wu, and German G. Creamer. Semantic frames to predict stock price movement. In *In Proc. of the 51th Annual Meeting of the Association for Computational Linguistics*, pp. 873–883, 2013.

[14] Frank Z. Xing, Erik Cambria, and Roy E. Welsh. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, Vol. 50, No. 1, pp. 49–73, 2018.

[15] Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *In Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1970–1979, 2018.

[16] 伊藤諒, 坂地泰紀, 和泉潔, 須田真太郎. 語の類義性・対義性を考慮したドメイン特化型辞書構築手法の提案と評価. 金融情報学研究会 (SIG-FIN), 2017.