DEIM Forum 2022 E31-1

株価上昇・下落ニュースの機械読解

鈴木　勢至† 　堤　 楽人† 　宇津呂武仁†

† 筑波大学大学院　システム情報工学研究群　知能機能システム学位プログラム
〒 305–8577 茨城県つくば市天王台 1 丁目 1 番 1 号

あらまし　本論文では株価の変動理由をニュースから抽出する手法を提案する．本論文では，BERT ベースの機械読解モデルを採用し，株価変動に関するニュース報道から株価変動の理由を抽出する．機械読解モデルの学習には読解事例を用いるが，一般に読解事例の生成には質問 Q を手動で作成する必要がある．本論文のアプローチでは読解事例の質問 Q に対して記事のタイトルを用いており，対象とした株価変動ニュースから株価変動理由を抽出する機械読解モデルを訓練するための読解事例を作成した．そして，作成した読解事例を用いて読解モデルの訓練と評価を行った．その結果，質問 Q を手動で作成する場合と同等の性能が確認できた．また，株価の下落理由を対象とした読解事例を本アプローチで作成し，同様にファインチューニングを行った結果，上昇理由と同等の性能が確認できた．本アプローチは読解事例の生成時に質問を手動で作成する手順を省略できる点において有望である．

キーワード　機械読解，BERT，株価ニュース，ファイナンス

1　はじめに

株価が大きく変動する理由として，投資者向けの情報である IR の発表や，実際にその企業で起きた出来事のニュース報道が挙げられる．これらの情報が発信されると，図 1 に示すように，その企業の株の取引き量を表す出来高が増加し，株価に大きな変動が起こる事例が多く存在する．このような株価の変動が起きた時に，ファイナンスに関するニュースサイトでは，図 3 のように株価の変動とその理由を報じることがある．本論文では，このように株価の変動とその理由を報じるニュースを「株価変動ニュース」，株価の変動を示す語彙を「変動ワード」，株価が変動した理由を「株価変動理由」と呼ぶ．株価変動ニュースは，どのような出来事によってどのように株価が変動するのかを知るには非常に有用な情報であるが，人手によって書かれているため，全ての株価変動について報じられているわけではない．銘柄によっては，株価が急激に変動した場合でも，株価変動ニュースにおいてその銘柄の株価変動が報じられない事例が存在する．そこで本論文では，株価変動ニュースにおいて株価変動理由が報じられない事例を想定して，ウェブ全体から当該企業の動向を収集して，株価変動理由に相当する動向を抽出するというアプローチをとる．そして，そのためのモデルを訓練するための訓練事例集合を構築するための準備研究として，株価変動ニュースに対して機械読解モデル [3], [7] を適用し，株価変動理由を抽出する手法を確立する．以上をふまえて，本論文では，株価変動ニュースから株価変動理由を抽出する機械読解モデル (図 2) を訓練するための機械読解データセットを作成した．そして，作成したデータセットを用いて読解モデルの訓練・評価を行った．

機械読解モデルによる機械読解タスクは，質問，および，その質問に対する回答が記述されている文書であるコンテ

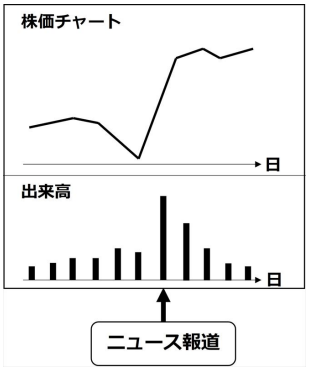


図 1　一日あたりの株価と出来高，ニュース報道の関係

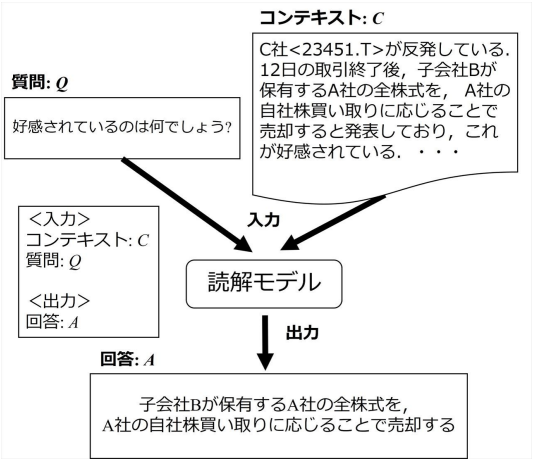


図 2　株価変動ニュースを対象とする読解モデルの枠組み

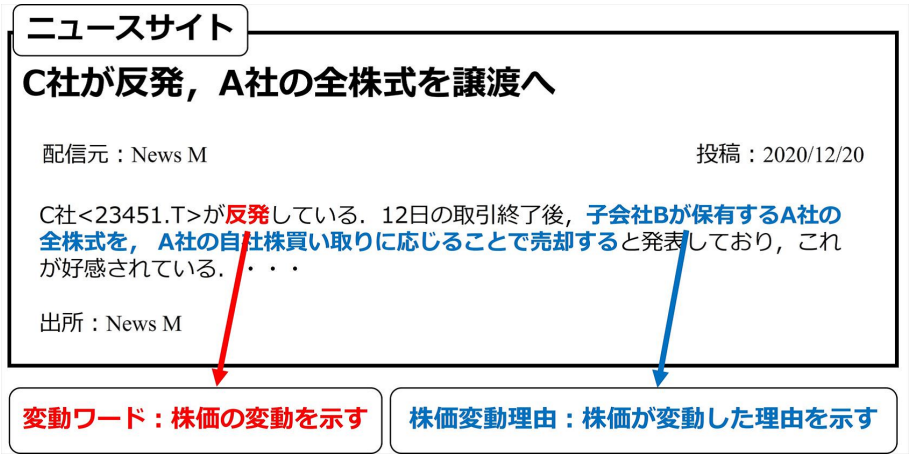
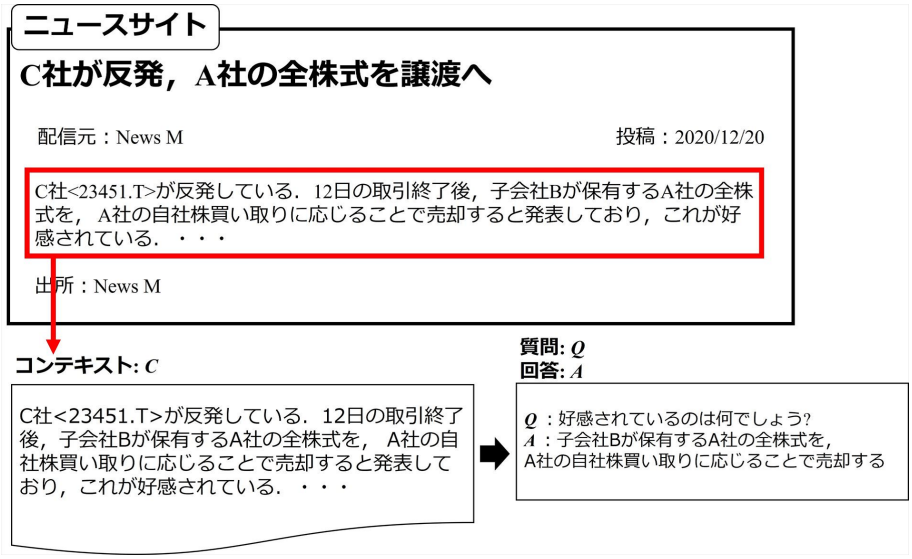
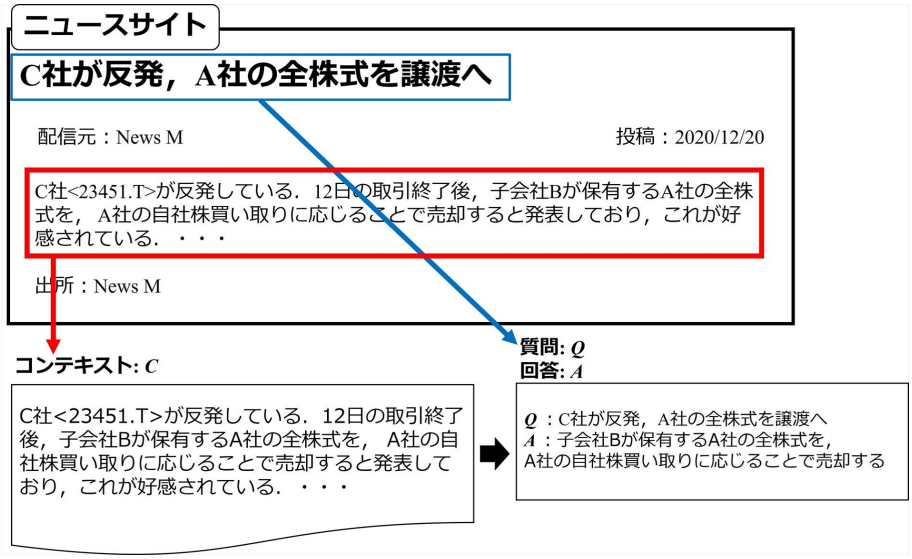


図 3　変動ワードと株価変動理由の例



(a) 質問 Q を手動で作成した場合



(b) 質問 Q として記事のタイトルを用いた場合

図 4　株価変動理由に関する機械読解事例の作成手順

表 1　「配信元=みんなの株式」の記事 100 件中のジャンルの内訳

ジャンル	概要	件数
株価変動ニュース	株価の変動とその理由に関するニュース	25
企業ニュース	新製品の発表等の企業に関するニュース	7
国内株式市場	国内株式市場全体の情報	24
外国株式市場	外国株式市場全体の情報	9
為替市場	為替情報	7
債券市場	債券の情報	2
金融市場	金・石油等の情報	2
個人投資家の予想	個人投資家向け情報	21
経済指標	市場に影響を与えるイベントの情報	3
合計		100

キストを入力として，コンテキスト中の回答部分を抽出するタスクとして定式化される．このタスクは，固有名詞や数量などの事実を回答対象とした事実型機械読解，および，物事のやり方や理由などの非事実を回答対象とした非事実型機械読解の二種類に大別される．事実型機械読解においては，人間による読解の性能を上回る性能を達成したことが知られている [1]．これに対して，本論文で対象とする株価変動理由の機械読解タスクは，非事実型機械読解に分類される．非事実型機械読解の一例として，ノウハウを対象とした機械読解の研究事例 [2] が知られている．この事例においては，ノウハウ機械読解モデル訓練用事例集合を用いた BERT [3] の fine-tuning により，一定のノウハウ機械読解性能が達成できることが報告されている．

2 「みんなの株式」による株価変動ニュース

本論文では，ファイナンスに関するニュースサイトとして `minkabu.jp`[2] を対象とし，そのうち，株価変動ニュースを多く含む配信元として，「配信元=みんなの株式」となるニュース記事 3,300 件 [3] を収集した．`minkabu.jp` においては，「配信元=みんなの株式」となるニュース記事約 26 万件 (2021 年 1 月時点) が掲載されている．次に，収集された 3,300 件から無作為に選定した 100 件におけるジャンルの内訳を表 1 に示す．この結果において「株価変動ニュース」は 25 件 (25%) であったため，「配信元=みんなの株式」となるニュース記事全 26 万件中，「株価変動ニュース」の推定件数は約 6 万 5,000 件である．以上より，「配信元=みんなの株式」となる記事を収集することにより，株価変動ニュースが一定規模収集可能であることから，株価変動理由機械読解事例作成の知識源としては十分な規模であると言える．

一般に株価の変動としては，「上昇」，「下落」の二種類の変動があるが，機械読解事例と読解モデルの条件変更実験ではこのうち，株価の「上昇」を対象として株価変動理由の機械読解事例を作成する．まず，あらかじめ，株価上昇において用いられる「上昇変動ワード」を調査し，表 2 に示す 11 種類以上の「上昇変動ワード」を含む株価変動ニュース 627 件を，全 3,300 件の中から無作為に選定した．[4]

表 2　株価変動理由機械読解事例 627 件中の上昇変動ワードの出現回数・割合 (%)

上昇変動ワード	回数	割合
反発	176	19.7
続伸	172	19.3
高値	115	12.9
カイ気配	87	9.7
大幅高	66	7.4
上昇	56	6.3
ストップ高	54	6.0
急伸	49	5.5
連騰	40	4.5
堅調	38	4.3
急騰	35	3.9
その他	5	0.5
合計	893	100

表 3　質問 Q の件数・割合 (%)

質問 Q	件数	割合
材料視されているのは何でしょう?	130	20.7
好感されているのは何でしょう?	122	19.5
寄与したのは何でしょう？	61	9.7
要因は何でしょう?	48	27.7
発表したのは何でしょう?	38	6.1
好調なのは何でしょう?	29	4.6
利益を押し上げたのは何でしょう？	25	4.0
期待されているのは何でしょう？	14	2.2
背景は何でしょう？	12	1.9
その他 (件数 10 以下)	148	23.6
合計	627	100

3 株価変動ニュースからの株価変動理由機械読解事例の作成

株価ニュース記事から株価変動原因の機械読解例を生成

1：英語版 Wikipedia の記事から作成した機械読解タスク用データセット SQuAD [7] が対象の場合 (`https://rajpurkar.github.io/SQuAD-explorer/`).

2：`https://minkabu.jp/`

3：2020 年 10 月 30 日 ～12 月 3 日の期間に配信された記事を収集した．

4：ただし，記事の形式上不適切な 14 記事はあらかじめ除外した．また，「反落」，「下落」，「続落」，「急落」，「売りに押され」，「安値」等の 13 種類の「下落変動ワード」を含む 35 記事についても，あらかじめ除外した．

する手順では，以下の二種類のアプローチを評価する．

(a) 図 4(a) の「質問 Q を手動で作成した場合」．この場合，記事から手動で生成した質問を，機械読解の質問 Q として使用する．

(b) 図 4(b) の「質問 Q として記事のタイトルを用いた場合」．この場合，記事のタイトルを，機械読解の質問 Q として使用する．

いずれの場合も，機械読解のコンテキスト C としては，記事全文を使用する．回答スパンを含む一文の中には，通常，株価上昇の「原因」を示す語彙（以下,「原因」語彙と呼ぶ）が出現する．これらの「原因」語彙について今回使用する読解事例には 62 種類の「原因」語彙が現れており，これらの一例を下記に示す．

「材料視」,「好感」,「期待」,「発表」,「株価を刺激」,「要因」,「引き上げ」,「追い風」

(a) および (b) の二種類のアプローチを比較すると，アプローチ (b) においては，質問 Q を記事タイトルから自動生成できる点において，アプローチ (a) よりも圧倒的に有利である．一方，アプローチ (a) の場合は，表 3 に示すように,「原因」語彙から質問文を手動で生成するという多大なコストを要する．[5]

なお，大規模機械読解 [1], [4] の設定において，質問 Q のみが与えられ，コンテキスト C の候補を大規模文書集合中から収集する場合には，収集したコンテキスト候補の中から「原因」語彙を特定し，特定した「原因」語彙に応じて質問 Q を自動的に切り替え，質問 Q とコンテキスト C の組に対して機械読解モデルを適用するアプローチが有望である．

4 評　　価

日本語 BERT [3] の実装として，TensorFlow のバージョン [6] と NICT BERT 日本語 Pre-trained モデル [7] [8] を事前学習モデルとして使用した．BERT モジュールを適用する前に，MeCab[9] と mecab-ipadic-NEologd 辞書 [10] を適用し，日本語テキストを形態素列に分割した．最後に，機械読解のための BERT の fine-tuning モジュール [11] を適用した．一連の手順では，まず，27,427 件の事実型機械読解訓練事例 [12] を用いて BERT の事前学習モデルを fine-tuning した後，前節で開発した株価変動原因の機械読解用訓練事例を用いて，さらに fine-tuning を行った．

評価として，5 分割交差検定によって，図 4(a) の「質問 Q を手動で作成した場合」，および，図 4(b) の「質問 Q として記事のタイトルを用いた場合」を比較した結果を図 5 に示す．また，記事の中には株価の変動理由になり得る文が複数存在する場合があるため，評価時の参照回答は複数設定している．この結果からわかるように，両者はほぼ同じ性能を達成している．このことから，両者がほぼ同じ性能を達成しているにもかかわらず,「質問 Q として記事のタイトルを用いた場合」には，記事から質問を生成するという手動の手順を避けることができる点において有益であることを示している．

表 4　株価変動理由機械読解事例 550 件中の下落変動ワードの出現回数・割合 (%)

下落変動ワード	回数	割合
嫌気	341	24.7
反落	230	16.7
続落	156	11.3
赤字	153	11.1
急落	104	7.5
減益	73	5.3
出尽くし感	50	3.6
転落	49	3.6
下落	32	2.3
その他	192	13.9
合計	1,380	100

5 「下落」の機械読解事例評価

株価の変動の「上昇」について実験を行ったが，株価の変動には「下落」も存在するため「下落」の機械読解事例を作成し，機械読解モデルの評価を行った.「下落」の機械読解事例の作成には,「上昇」と同様に minkabu.jp を対象とした,「配信元＝みんかぶ」の記事を 15,300 件収集した．その中から表 4 の「下落変動ワード」と「下落原因語彙」[13] を含む「下落」の株価変動ニュースを選定し,「下落」の機械読解事例を 550 件作成した．機械読解事例は質問 Q を記事のタイトルとして，コンテキスト C を記事全文としている．　評価に用いた機械読解モデルは「上昇」と同じものであり，評価も 5 分割交差検定によって行った．ただし，評価時の参照回答 A は単一としている．結果は図 6 に示す．この結果からわかるように,「上昇」も「下落」も同程度の性能を達成している．このことから，本論文の手法は「上昇」と「下落」両方の株価の変動に対応することができることが示された．

6 関 連 研 究

酒井ら [8] は日経平均株価の市場分析記事から市場分析コメントを生成することを目的として，因果関係と補足情報の抽出手法の提案と評価を行った．本論文では，酒

5 : 第三のアプローチとして，

例 :「A 社の株価が変動した理由はなんでしょう?」

のような一般的な質問文 Q を共通に一つ使用するアプローチも評価したが，よい性能は得られなかった．

6 : https://github.com/google-research/bert

7 : 日本語 Wikipedia を対象に情報通信研究機構 データ駆動知能システム研究センターで事前学習を行った BERT モデル．

8 : https://alaginrc.nict.go.jp/nict-bert/index.html

9 : http://taku910.github.io/mecab/ (in Japanese)

10 : https://github.com/neologd/mecab-ipadic-neologd 辞書データ 2021/11/08 時点

11 : run_squad.py, エポック数 2，バッチサイズ 8，学習率 0.00003

12 : http://www.cl.ecei.tohoku.ac.jp/rcqa/

13 : 下落原因語彙の一例 :「発表」,「嫌気されている」,「影響」,「見通し」,「営業損益」,「響いた」,「下方修正」,「要因」

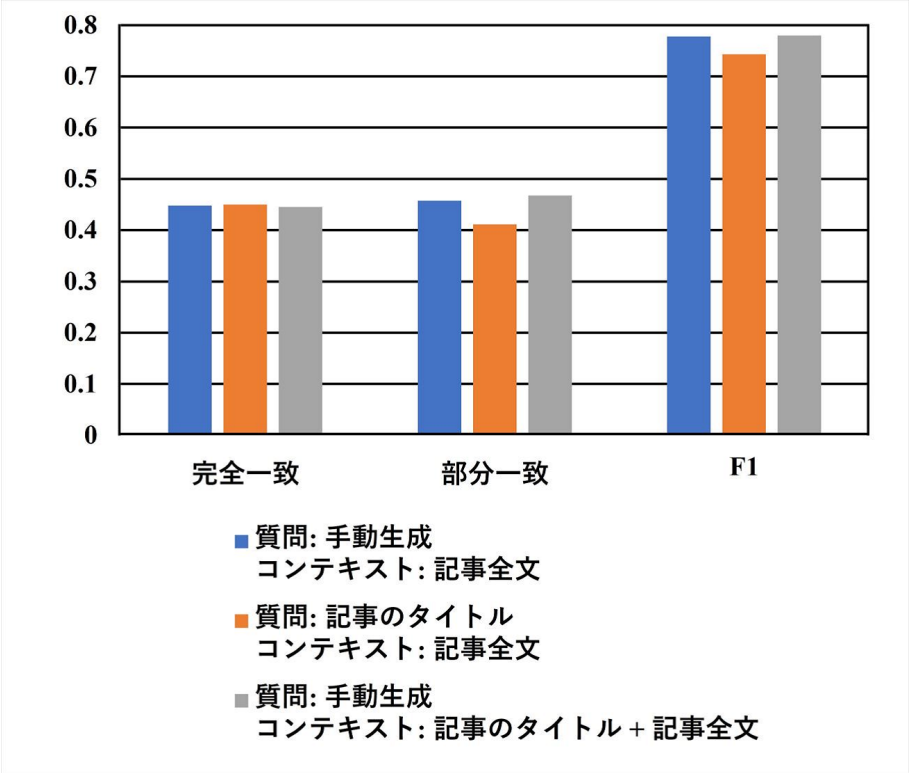


図 5　株価「上昇」ニュース機械読解の評価結果

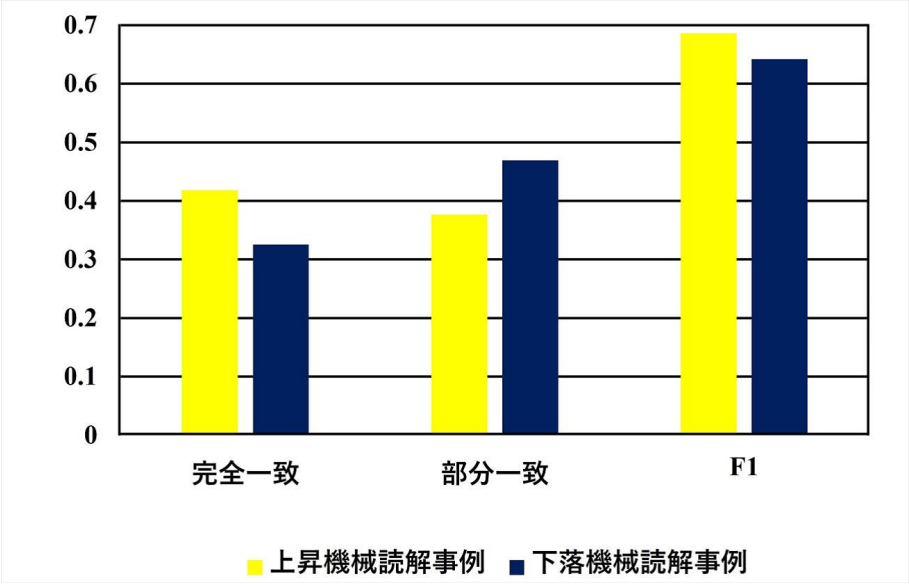


図 6　機械読解事例「上昇」「下落」比較結果

井ら [8] と比較して，日経平均株価という市場全体の分析コメントではなく，個々の銘柄の株価変動の原因を機械読解の抽出手法に基づいて読み取ることに重点を置いている． 一方で金融分野を対象としたタスクも存在する．Maia ら [5] は”Financial Opinion Mining and Question Answering”（FiQA）[14] を提案した．FiQA のタスク 2 ”Opinion-based QA over financial data” は金融分野における意見に関する質問応答タスクとなっている．また，Mariko ら [6] は”Financial Document Causality Detection Shared Task”(FinCausal 2020) を提案した．これは金融分野における原因と影響を抽出するタスクである．一方，本論文では，株価の上昇・下落理由の回答タスクを対象としているため，金融分野における意見，あるいは，原因と影響を対象とするタスクとは異なるタスクとなっている．

7 おわりに

本論文では，BERT [3] ベースの機械読解モデル [7] を採用し，株価変動に関するニュース報道から株価変動の理由を抽出するというアプローチを提案した．本論文のアプローチでは，機械読解の質問 Q に対して記事のタイトルを用いるアプローチにより，記事から質問を手動で生成する手順を省略できる点において有望である．今後の課題としては，大規模機械読解の枠組み [1], [4] においてコンテキスト C の候補を自動収集することが挙げられる．また，最新のニュースを自動収集し，それらから株価変動理由を抽出し，今後株価が上昇するのか下落するのかといった株価の変動自体を，予測するシステムの開発も視野に入れている．

文　献

[1] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading Wikipedia to answer open-domain questions. In *Proc. 55th ACL*, pp. 1870–1879, 2017.

[2] 陳騰揚, 前田竜冶, 李宏宇, 銭澤長, 宇津呂武仁, 河田容英. ウェブ上のコラムページを情報源とする回答不可能なノウハウ質問応答事例の作成. 言語処理学会第 26 回年次大会論文集, pp. 315–318, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[4] K. Lee, M.-W. Chang, and K. Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proc. 57th ACL*, pp. 6086–6096, 2019.

[5] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahu. WWW 18 open challenge: Financial opinion mining and question answering. In *Proc. WWW*, pp. 1941–1942, 2018.

[6] D. Mariko, H. Abi-Akl, E. Labidurie, S. Durfort, H. De Mazancourt, and M. El-Haj. The financial document causality detection shared task (FinCausal 2020). In *Proc. 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pp. 23–32, 2020.

[7] R. Pranav, Z. Jian, L. Konstantin, and L. Percy. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, pp. 2383–2392, 2016.

[8] 酒井浩之, 坂地泰紀, 和泉潔, 松井藤五郎, 入江圭太郎. 学習データ自動生成による市況分析コメント作成のための要因文と補完情報の抽出. 第 34 回人工知能学会全国大会論文集, 2020.

14 : https://sites.google.com/view/fiqa/home