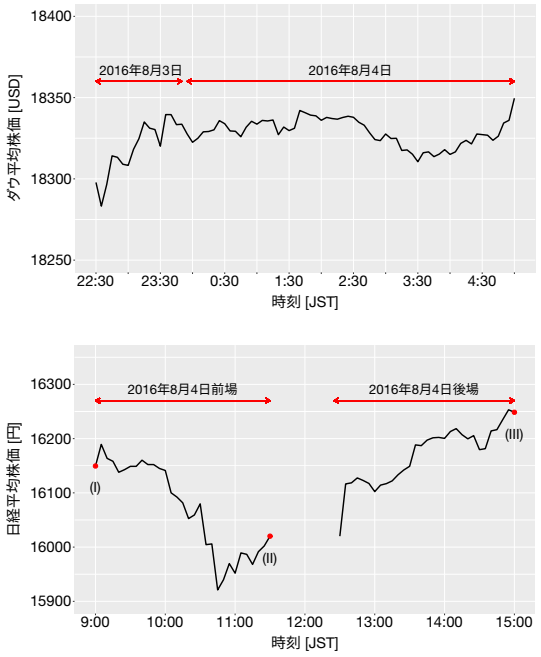
時系列数値データにおける変化要因の記述

青木 竜哉[†,§] 宮澤 彬[‡,¶,§] 青木 花純[◇,§] 五島 圭一[*,§] 小林 一郎[◇,§] 高村 大也[†,§] 宮尾 祐介[‡,¶,§]
[†]東京工業大学 [‡]総合研究大学院大学 [◇]お茶の水女子大学 [*]日本銀行 [¶]国立情報学研究所 [§]産総研
aoki@lr.pi.titech.ac.jp, keiichi.goshima@boj.or.jp, takamura@pi.titech.ac.jp,
{miyazawa-a, yusuke}@nii.ac.jp, {g1120501, koba}@is.ocha.ac.jp

1 はじめに

数値データの動きを解釈し，テキストでの説明を生成する技術は，大量の数値データを理解し有効利用するために必要不可欠である．テキストによる説明は，数値の変化を記述するだけでなく，しばしばその変化要因に言及する．図 1 の概況テキストは日経平均株価指数に関するものであるが，その変化要因として米国株が上昇傾向にあり，円高が止まりつつあることを述べている．このような変化要因の記述は，データに対する理解を促進させ，データのさらなる有効利用につながると考えられる．そこで，本研究では数値変化の記述に加え，「米株高を好感」のような変化要因の記述の生成を目的とする．

ただし，変化要因の候補は一般に複数存在し，それらのうちのどれが変化要因と考えられるかは明らかでない．日経平均株価指数の変化においても，図 1 に示したような米国株に加え，ドル円外国為替，アジア株，原油価格などを含む多くの変化要因候補が考えられる．これらの変化要因候補の中から適切なものを選び，それについて記述できる必要がある．そこで，主な記述対象である指標 (ここでは日経平均株価) に加え，変化要因候補である複数の金融指標の時系列データを入力として与え， 変化要因に関する記述を含む概況テキストを生成する課題に取り組む．ただし，出力が制御できるように，数値変化の記述だけを生成するのか，変化要因の記述も合わせて生成するのかは，追加入力として与えることとする．



	時刻	概況テキスト
(I)	09:07	日経平均、反発で始まる　米株高を好感
(II)	11:34	日経平均、1 万 6000 円下回る　下げ幅 100 円超える
(III)	15:02	日経平均大引け、3 日ぶり反発　171 円高、円高一服で買い戻し

図 1: ダウ平均株価の値動き（上）と日経平均の値動き（下），およびそれを説明する概況テキスト（表）．

2 関連研究

2.1 データからのテキスト生成

数値データなどに対しそれを説明するテキストを生成する研究課題の歴史は長く，かつては何を言うかを決定するテキストプランニング，どう言うかを決定するマイクロプランニング，実際にテキストとして出力する表層化の三段階に分けて論じられ，天気予報コメントの生成などを題材に技術開発が進められた [2]．しかし，最近では end-to-end の学習を行うエンコーダ・デコーダモデルが高い性能を持つことがわかってきた．エンコーダ・デコーダモデル [7] を用いてデータからテキスト生成をする試みとしては，属性情報から製品レビューを生成する研究 [1]，構造化データから人物説明テキストを生成する研究 [4]，気象データから天気予報コメントを生成する研究 [9]，時系列数値データである株価から市況コメントを生成する研究 [6] などがある．これらのうち最後の研究は，我々の目的と最も関連が深いので，2.2 節にて詳しく説明する．

* 本稿に示されている意見は，筆者たち個人に属し，日本銀行の公式見解を示すものではない．また，あり得べき誤りはすべて筆者たち個人に属する．

2.2 時系列数値データを用いた概況生成

Murakami ら [6] は，エンコーダ・デコーダモデルに基づいて，株価データと株価の動きに言及しているニュースヘッドラインを例に，時系列数値データから概況テキストを生成するモデルを開発した．

Murakami らの手法では，入力として，日経平均株価の短期的な変動を捉えるために 5 分足で収集した 1 取引日分の株価データ $\boldsymbol{x}_{\text{short}} = (x_{\text{short},1}, x_{\text{short},2}, \ldots, x_{\text{short},N})$ を，また長期的な変動を捉えるために 1 取引日ごとに収集した 7 取引日分の株価の終値データ $\boldsymbol{x}_{\text{long}} = (x_{\text{long},1}, x_{\text{long},2}, \ldots, x_{\text{long},M})$ をそれぞれ用いている．そして，これらの時系列数値データを数値ベクトルで表現するために，短期的変動を捉えるためのエンコーダと，長期的変動を捉えるためのエンコーダの 2 種類を用意し，それぞれ $\boldsymbol{x}_{\text{short}}$ および $\boldsymbol{x}_{\text{long}}$ に対して前処理を加えたデータをエンコーダへの入力としている．デコーダへは，2 種類のエンコーダから出力された数値ベクトルと前処理した数値データを結合した *multi-level representation* ベクトルを入力する．加えて，ニュースヘッドラインの配信時間帯を埋め込みベクトルとして各時刻の状態に追加入力することで，RNNLM モデルを用いた日経平均株価の数値変化を概況するテキスト（単語列）を推定している．

3 提案手法

3.1 エンコーダ・デコーダモデル

本研究では，Murakami らのモデルをベースにし，日経平均株価データをエンコードし，その情報をデコーダの初期値として使用するエンコーダ・デコーダモデルを用いる．これに加えて，ダウ平均株価のような外部データも参照する注意機構を組み込むことによって，変化要因の記述も可能な概況テキスト生成モデルを提案する．提案モデルの概要を図 2 に示す．

まず，2.2 節で導入した $\boldsymbol{x}_{\text{short}}$ および $\boldsymbol{x}_{\text{long}}$ に対して，それぞれ Murakami らの研究で最も有効であるとされた 2 種類の前処理を適用する[*1]．1 つ目は，平均値と標準偏差を用いて標準化する手法であり，この処理を $\boldsymbol{x}_{\text{short}}$ および $\boldsymbol{x}_{\text{long}}$ に適用して得られた数値データをそれぞれ，$\boldsymbol{x}^{\text{std}}_{\text{short}}$ および $\boldsymbol{x}^{\text{std}}_{\text{long}}$ と表す．2 つ目は，それぞれの入力データに対して前取引日の終値からの差分を計算する手法であり，この処理で得られた数値データをそれぞれ，$\boldsymbol{x}^{\text{move}}_{\text{short}}$ および $\boldsymbol{x}^{\text{move}}_{\text{long}}$ と表す．

短期的変動を捉えるためのベクトルを $\boldsymbol{l}_{\text{short}} = \left[\boldsymbol{x}^{\text{std}}_{\text{short}}; \boldsymbol{x}^{\text{move}}_{\text{short}}\right]$，$\boldsymbol{l}_{\text{long}} = \left[\boldsymbol{x}^{\text{std}}_{\text{long}}; \boldsymbol{x}^{\text{move}}_{\text{long}}\right]$ とすると，デコーダの初期値として与える隠れ状態 $\boldsymbol{s}_0$ は，

$$\boldsymbol{s}_0 = W\left(\left[\boldsymbol{l}_{\text{short}}; \boldsymbol{l}_{\text{long}}; \boldsymbol{h}_{\text{short}}; \boldsymbol{h}_{\text{long}}\right]\right) + \boldsymbol{b} \tag{1}$$

となる．ここで $\boldsymbol{h}_{\text{short}}$ および $\boldsymbol{h}_{\text{long}}$ は，それぞれ $\boldsymbol{l}_{\text{short}}$，$\boldsymbol{l}_{\text{long}}$ を対応するエンコーダ $\text{Encoder}_{\text{short}}$ と $\text{Encoder}_{\text{long}}$ を用いて，$\boldsymbol{h}_{\text{short}} = \text{Encoder}_{\text{short}}(\boldsymbol{l}_{\text{short}})$，$\boldsymbol{h}_{\text{long}} = \text{Encoder}_{\text{long}}\left(\boldsymbol{l}_{\text{long}}\right)$ のようにエンコードして得られるベクトルである．また $\boldsymbol{b}$ はバイアス項である．時刻 t におけるデコーダの隠れ層の状態 $\boldsymbol{s}_t$ は，時間帯情報埋め込みベクトル $\boldsymbol{T}$，直前の単語埋め込み $\boldsymbol{w}_{t-1}$，直前の隠れ層の状態 $\boldsymbol{s}_{t-1}$ を用いて次のように計算される:

$$\boldsymbol{s}_t = \text{LSTM}(\boldsymbol{T}, \boldsymbol{w}_{t-1}, \boldsymbol{s}_{t-1}). \tag{2}$$

最終的に時刻 t での各単語の出力確率分布は，デコーダの隠れ層の状態 $\boldsymbol{s}_t$ を用いて，

$$\text{softmax}(W_s \boldsymbol{s}_t) \tag{3}$$

と表される．ここで W_s は重みを表す．

3.2 外部データを組み込んだ注意機構付きエンコーダ・デコーダモデル

提案手法では，3.1 節で説明したエンコーダ・デコーダに対して，ダウ平均株価や円ドル為替のような外部指標を組み込む．具体的には，使用する外部データごとに新たにエンコーダを用意し，これにより各外部データを数値ベクトルへ変換したのち，デコーダの隠れ状態と新たにエンコードした複数の外部データとの対応付けを注意機構によって実現する．これにより，デコーダの各ステップにおいて，外部データを参照しつつ出力単語を予測するモデルを構築した．

まず，日経平均以外の株価指数や外国為替といった外部データを n 種類用意し，これを $\boldsymbol{x}^{\text{ext}\,1}_{\text{long}}, \boldsymbol{x}^{\text{ext}\,2}_{\text{long}}, \ldots, \boldsymbol{x}^{\text{ext}\,n}_{\text{long}}$ とする．このとき各 $\boldsymbol{x}^{\text{ext}\,i}_{\text{long}}$ の数値データには，それぞれの外部データにおける 7 取引日分の終値を使用する．これらに，日経平均と同様の前処理を行なった数値データを $\hat{\boldsymbol{x}}^{\text{ext}\,1}_{\text{long}}, \hat{\boldsymbol{x}}^{\text{ext}\,2}_{\text{long}}, \ldots, \hat{\boldsymbol{x}}^{\text{ext}\,n}_{\text{long}}$ とし，式 (4) によって，前処理済み数値データをベクトル $\boldsymbol{v}_i$ へ変換する:

$$\boldsymbol{v}_i = \text{Encoder}^{\text{ext}\,i}_{\text{long}}\left(\hat{\boldsymbol{x}}^{\text{ext}\,i}_{\text{long}}\right). \tag{4}$$

ただし，$\text{Encoder}^{\text{ext}\,i}_{\text{long}}$ は，それぞれの外部データ $\hat{\boldsymbol{x}}^{\text{ext}\,i}_{\text{long}}$ ごとに新たに用意したエンコーダを示す．そして，$\boldsymbol{h}^{\text{ext}}_{\text{long}} = [\boldsymbol{v}_1; \ldots; \boldsymbol{v}_n]$ とすると，

$$\boldsymbol{s}_0 = W\left(\left[\boldsymbol{l}_{\text{short}}; \boldsymbol{l}_{\text{long}}; \boldsymbol{h}_{\text{short}}; \boldsymbol{h}_{\text{long}}; \boldsymbol{h}^{\text{ext}}_{\text{long}}\right]\right) + \boldsymbol{b} \tag{5}$$

としてデコーダの隠れ状態を初期化する．

時刻 t において，デコーダの隠れ状態を $\boldsymbol{s}_t$ とすると，注意機構によってそれぞれの外部データで重み付けさ

[*1] 詳細は Murakami ら [6] の 3.1 節を参考にされたい．

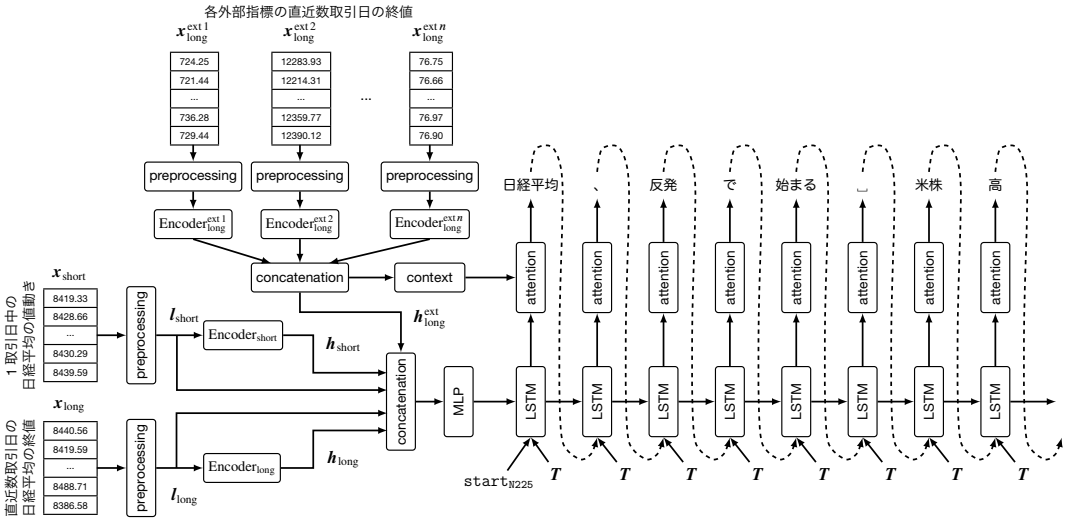


図 2: 提案モデルの概要

れた隠れ状態 $\tilde{\boldsymbol{s}_t}$ は，次の式 (6) によって表せる:

$$\tilde{\boldsymbol{s}_t} = \tanh\left(W_c[\boldsymbol{c}_t; \boldsymbol{s}_t]\right). \tag{6}$$

ただし W_c は重みのパラメータである．最終的に，式 (3) における $\boldsymbol{s}_t$ を $\tilde{\boldsymbol{s}_t}$ に置き換えて，出力単語を予測する．また $\boldsymbol{c}_t$ は，エンコードされた外部データ $\boldsymbol{v}_i$ とそれに対応する重み a_{ti} を用いて以下のように計算される:

$$\boldsymbol{c}_t = \frac{1}{n}\sum_{i=1}^{n} a_{ti}\boldsymbol{v}_i. \tag{7}$$

デコーダの時刻 t で，エンコードされた外部データ $\boldsymbol{v} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ に対する重み $\boldsymbol{a}_t = (a_{t1}, \ldots, a_{tn})$ は，$\boldsymbol{v}$ とデコーダの隠れ状態 $\boldsymbol{s}_t$ を用いて

$$a_{ti} = \frac{\exp\left(\text{score}\left(\boldsymbol{s}_t, \boldsymbol{v}_i\right)\right)}{\sum_{j=1}^{n}\exp\left(\text{score}\left(\boldsymbol{s}_t, \boldsymbol{v}_j\right)\right)} \tag{8}$$

により求める．重み付けのための関数 score には Luong ら [5] の手法における concat モデルを使用した．

3.3 概況テキストの種類を表すタグの導入

日経平均などの株価指数の変動を説明する際には，日経平均そのものの値動きだけではなく，ダウ平均株価のような外部データの変動も合わせて伝える方が，情報量の観点から好ましい．実際に概況テキストの中には，図 1 中の概況テキスト (I) や (III) のように，「米株高」や「円高一服」というような，外国為替や他の株価指数などに言及しているものが見られる．

そこで，本研究では，概況テキストを生成する際に，生成するテキストが，日経平均の変動のみに言及するのか，または外部データと関連付けて言及するのかといった，生成するべき概況テキストの種類をあらかじめモデルに与えた上で，その種類に応じた概況テキスト生成を行う．本研究では，図 1 中の概況テキスト (I), (III) のような，米国株や為替など外部のデータを参照している場合と，そうでない場合についてそれぞれ異なる記号 $\texttt{start}_{\text{ext}}$，$\texttt{start}_{\text{N225}}$ を，デコーダの開始記号として用いる．Yamagishi ら [8] は，学習コーパスに出力するべき文の態情報を組み込むことで，エンコーダ・デコーダモデルによる出力文の態制御を実現している．本研究においては，何に言及すべきかを開始記号として与えることで，出力文の制御を試みる．各ヘッドラインに対して，これらのタグの適切な方を先頭に付与し，その文のタイプを表す開始記号として機能させることを目的としている．

ただし，複数の外部データ (すなわち日経平均の変化要因候補) のうちどれに言及するかについては，モデルが自動的に選び，変化要因を記述する．

4 実験

4.1 実験設定

本研究では，時系列データとして，Thomson Reuters DataScope Select[*2]から 2011 年 1 月から 2016 年 9 月までの期間における，日経平均株価指数やダウ平均株価を含む 8 種類[*3]の時系列データを収集し，実験に使用した．概況テキストとして，日経 QUICK ニュース社が提供している日経平均株価に言及しているニュース記事のヘッドラインを使用した．実験に使用したデータの概要を表 1 に示す．2011 年から 2014 年の期間のデータを学習データ，2015 年のデータを開発データ，2016 年のデータを評価データとして使用した．

[*2] https://hosted.datascope.reuters.com/DataScope/

[*3] 日経平均株価，東証株価指数，ダウ平均株価，S&P500，上海総合指数，FTSE100 種総合株価指数，USD/JPY，USD/EUR

表 1: 使用したデータの統計値

期間	要因記述		合計
	なし	あり	
2011/01/01 – 2014/12/31	4,831	2,427	7,258
2015/01/01 – 2015/12/31	1,670	815	2,485
2016/01/01 – 2016/09/30	1,349	781	2,130

表 2: 評価データにおける BLEU (%)

手法	要因記述		全体
	なし	あり	
Murakami ら [6]	18.44	11.68	15.81
Murakami ら [6] + タグ	18.33	12.43	16.05
Murakami ら [6] + タグ (512 次元)	18.80	13.86	16.87
提案手法	21.34†	16.59†	19.45†

ヘッドラインには，あらかじめ前処理を加えて，不要な語の除去を行なった上で，MeCab[*4]による形態素解析を IPA 辞書を用いて行った．このとき「日経平均」や「前引け」のような用語を新たに辞書に追加した．

式 (1) における，株価等の数値データを数値ベクトルへ変換する際の各エンコーダには，Murakami らの研究で最も性能が高かった MLP（多層パーセプトロン）を使用した．外部データに対するエンコーダの隠れ状態の次元は 64 とし，その他のパラメータは Murakami らの研究と同一の値を使用した．モデルの最適化手法は Adam を使用し，学習率は 10^{-5} を使用した．評価指標には，BLEU を採用し，モデルによって生成された文と，その時刻に書かれた実際の概況テキストとの一致度を測る．学習回数については，開発データで計算された BLEU が 6 回連続で下がった場合に学習をストップする．そして，開発データ内で BLEU が最大となったモデルを使用して，評価データに対し評価を行う．

評価実験では，Murakami らの手法，Murakami らの手法に対して 3.3 節で導入したタグを使用した手法，Murakami らの手法において，エンコーダの隠れ層の大きさを 2 倍（512 次元）し，数値データの表現力を高めた上でタグを使用した手法と比較する．

4.2 結果

表 2 の実験結果[*5]によると，Murakami らのモデルに 3.3 節のタグを導入したり隠れ層を大きくすることでわずかな BLEU の向上が見られるものの，提案モデルでは大幅に BLEU を向上させることがわかる．要因記述なしのデータにおいても BLEU が向上していることから，外部指標が日経平均株価の概況を生成する際の手がかりとなっていることが考えられる．

表 3: 生成されたヘッドラインの例

正解文	日経平均、続落で始まる　欧米株安や円高進行で
提案手法	日経平均、続落で始まる　米株安で、円高が重荷
従来手法	日経平均、反落で始まる　米株高や円安で

表 3 に，Murakami らの従来手法と本研究の提案手法によるヘッドライン生成例を示す．従来手法では，与えられた時刻における外部指標の値動きに関わらず，学習された言語モデルに基づいて概況テキストを生成していた．提案手法では，概況テキストの生成時に，その時刻における日経平均株価の値動きに加えて，米国株をはじめとする外部指標も併せて参照を可能にするモデルであるため，外部指標の変動が要因で日経平均株価が変動していた場合において，数値変化の記述に加え，「米株安で、円高が重荷」のような変化要因の記述が正確に生成される．

5 結論

本研究では，ダウ平均株価や外国為替などの外部指数の変動も参照しつつ，日経平均株価指数についての概況を生成するモデルを提案した．評価実験では，外部データの値動きの情報を考慮することで，概況生成の精度が有意に向上することが示された．

謝辞 この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務，および JST さきがけ JPMJPR1655 の支援の結果得られたものです．

参考文献

[1] Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. Learning to generate product reviews from attributes. In *EACL*, pp. 623–632, 2017.
[2] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *arXiv: 1703.09902*, 2017.
[3] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *EMNLP*, pp. 388–395, 2004.
[4] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *EMNLP*, pp. 1203–1213, 2016.
[5] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pp. 1412–1421, 2015.
[6] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. Learning to generate market comments from stock prices. In *ACL*, pp. 1374–1384, 2017.
[7] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.
[8] Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. Controlling the voice of a sentence in japanese-to-english neural machine translation. In *WAT*, pp. 203–210, 2016.
[9] 村上聡一朗, 笹野遼平, 高村大也, 奥村学. 数値予報マップからの天気予報コメントの自動生成. 言語処理学会年次大会, pp. 1121–1124, 2017.

[*4] http://taku910.github.io/mecab/

[*5] † は，有意水準を 5% とした時，Murakami ら [6] + タグ (512 次元) との差異が統計的に有意であることを示す．検定には，反復回数を 1,000 回，サンプルサイズを 100 としたときの paired bootstrap resampling[3] を用いた．