時系列データの概要文生成におけるトピックラベルによる内容制御

青木 花純[◇,§] 宮澤 彬[‡,§] 青木 竜哉[†,§] 能地 宏[§] 小林 一郎[◇,§] 高村 大也[§,†] 宮尾 祐介[¶,§]
[◇]お茶の水女子大学 [‡]総合研究大学院大学 [¶]東京大学 [†]東京工業大学 [§]産業技術総合研究所
{kasumi.aoki,koba}@is.ocha.ac.jp, aoki@lr.pi.titech.ac.jp,
hiroshi.noji@aist.go.jp, takamura@pi.titech.ac.jp, miyazawa-a@nii.ac.jp,
yusuke@is.s.u-tokyo.ac.jp

1 はじめに

画像や動画などの視覚情報や脳活動データ，金融関連データなど多岐にわたる大量の時系列数値データが利用可能となり，様々な形で活用されている．このようなデータの内，一部の専門的データの動向（内容）を解釈するためには専門知識が必要であり，専門家による動向（内容）説明文をデータに付与し，人の理解を助ける事が多く行われている．専門家による動向説明文の作成は時間がかかるため，説明文を自動生成する技術が求められているが，適切な説明文は読み手によって異なるため，生成の際に内容を制御するとともに，必要に応じて逐次的に文を生成する必要があると考えられる．例えば，日経平均株価に関して説明している市況ニュース記事は，日経平均株価ではなく米国の株価動向を知りたい読者や，このニュースに合わせてドルの為替レートの情報を知りたい読者にとっては不十分である．

そこで本研究では日経平均株価など複数の経済市況数値データを対象に，時系列データの動向概要文自動生成における生成文制御に取り組む．具体的には，文の生成内容を示すトピックラベルを必要に応じて入力として与える事で生成文の内容を制御し，逐次的に読み手にとって適切な文章を生成する．

2 関連研究

数値データなど非言語データに対して，データ内容を説明するテキストを自動生成する研究は以前より盛んに行われている．近年は，様々なデータを題材に，エンコーダ・デコーダモデルを用いて，end-to-end の学習を行うことで，高い生成性能を発揮している．具体的には，調理動作を含む動画情報を用いたキャプション生成 (漆原・小林, 2018) や気象データを用いた天気予報コメントの生成 (村上他, 2017)，株価などの経済市況データを用いた市況コメントの生成 (Murakami et al., 2017; Aoki et al., 2018) などがある．またテキスト生成においては，生成文の制御に関する研究も行われており，具体的には，文の極性の制御 (Peng et al., 2018) や storyline（物語の内容を表すキーワードの列）を用いた内容制御 (Yao et al., 2018) などがあげられる．

前述した研究のうち，複数の経済市況データを用いた市況コメント生成 (Aoki et al., 2018) および storyline を用いた物語生成 (Yao et al., 2018) は，本研究と関連が深いため，それぞれ 2.1, 2.2 節にて詳しく説明する．

2.1 時系列数値データの概要テキスト生成

Aoki et al. (2018) は，Murakami et al. (2017) が提案したエンコーダ・デコーダモデルを基に，複数の市況データとニュースヘッドラインを用いて，日経平均株価などの動向を示す概況テキストを生成するモデルを提案した．彼らの手法では，まず入力である各銘柄 i ($i = 1, \ldots, N$) の 1 取引日分のデータ $\boldsymbol{x}^i_{\text{short}}$ および 7 取引日分の終値データ $\boldsymbol{x}^i_{\text{long}}$ に対して平均値と標準偏差を用いて標準化を行うなど前処理を行い，短期変動，長期変動それぞれに対するベクトル $\boldsymbol{l}^i_{\text{short}}$，$\boldsymbol{l}^i_{\text{long}}$ を得る．また，これらのベクトルに対してそれぞれ MLP を用いて隠れ状態 $\boldsymbol{h}^i_{\text{short}}$，$\boldsymbol{h}^i_{\text{long}}$ を計算する．以上のベクトルや隠れ状態を用いて，市況データをエンコードした隠れ状態 $\boldsymbol{m}$ を得る：

$$\boldsymbol{m}^i = \boldsymbol{W}_m \left([\boldsymbol{l}^i_{\text{short}} ; \boldsymbol{l}^i_{\text{long}} ; \boldsymbol{h}^i_{\text{short}} ; \boldsymbol{h}^i_{\text{long}}] \right) + \boldsymbol{b}^i_m. \quad (1)$$

デコーダの入力には，前述した隠れ状態 $\boldsymbol{m}^i$ に加え，ニュースヘッドラインの配信時刻を埋め込んだ時間ベクトル $\boldsymbol{T}$ を各時点の隠れ状態に追加入力することで，生成文の制御を行っている．このモデルによって，入力された時間情報に基づき，データの動向を示す適切な文を生成することが可能だが，短い文を対象にしており，単語やフレーズなどを用いた具体的な内容の制御は行っていない．

2.2 生成テキストの制御

Yao et al. (2018) は，物語の内容を示すキーワード列である storyline を用いて物語生成における内容制御に取り組んでいる．Yao et al. の目的は多様な物語生成であり，物語の内容を title と storyline によって制御可能

なモデルを提案することで，インタラクティブに文生成を行うことを可能にしている．また，単語やフレーズから構成される storyline や title はテキスト表現であるため，ユーザにとって操作がしやすい．モデルの学習の際に使用する storyline は，自動キーワード抽出手法である RAKE (Rose et al., 2010) を用いて生成しているため，制御因子を人手で付与する必要がない．また，title と storyline のエンコードおよびテキスト生成には RNNLM を用いているため，end-to-end で学習が可能である．しかし，用いられている storyline は 1 単語もしくは 1 フレーズであり，複数の内容を表すような文生成を行うことは出来ない．また，title 以外に外部データを参照していないため，生成されたテキストは多様性や一貫性や文法，生成文の面白さの観点において評価されており，入力として与えられる市況データの動向内容を適切に記述する必要があるという本研究とは目的が異なる．ただし，単語やキーフレーズによって構成された因子を用いて内容を制御し，インタラクティブにテキスト生成を行うという点は本研究の目的と関連が深いといえる．

3 提案手法

Aoki et al. (2018) のモデルをベースとし，日経平均株価など複数市況データの概要テキスト生成において，テキストの内容を示すトピックラベルを用いたエンコーダ・デコーダモデルによりインタラクティブに概要テキストを生成する手法を提案する．概要図を図 1, 2 に示す．時系列データのエンコード手法については 2.1 節を参照されたい．

3.1 逐次的な文生成

前文の内容を考慮した逐次的な文生成を行うため，第 k 文の生成では，2.1 節で説明した市況データをまとめたもの $\boldsymbol{m} = [\boldsymbol{m}^1; \ldots; \boldsymbol{m}^N]$ に加え，前文のトークン列を LSTM でエンコードして得られた隠れ状態 $\boldsymbol{u}^{k-1}$ を以下の形でデコーダの初期状態として与える：

$$\boldsymbol{s}_0^k = \boldsymbol{W}_s^k \left(\left[\boldsymbol{m}; \boldsymbol{u}^{k-1} \right] \right) + \boldsymbol{b}_s^k. \tag{2}$$

3.2 トピック情報を用いたテキスト生成

追加情報を考慮したテキスト生成手法として，Aoki et al. (2018) は複数の市況データと時間帯情報 $\boldsymbol{T}$ を用いる手法，Yao et al. (2018) は内容を示す storyline を用いる手法を使用している．

これらを踏まえ，本研究ではデコード時の各時点の状態にトピック情報 $\boldsymbol{I}^k$ の付与を行い，トピック情報を考慮した概況テキスト生成を行う．トピックラベルは，

表 1: トピックラベルの構成要素の一覧．

〔対象〕日経平均, 日個別株, 日業種株, 東証一部, 東証二部, TOPIX, 米全体, 米個別株, 米業種株, ダウ 30, 香港全体, その他の国全体, その他の国個別株, その他の国業種株, 円全体, 円ドル, 円ユーロ, 円豪ドル, 円その他, ドル, 豪ドル, ユーロ, 香港ドル, 日債, 日債長期 (5&10), 日債 (2&3), 米債, 米債長期 (5&10), 米債短期 (2&3), TIBOR, 円金利, 経済指標, イベント, 材料なし, 投資家, その他, 買い入れオペ, 要人発言
〔現物・先物〕現物, 先物, その他
〔売買・動き〕売買, 動き, その他

〔対象物〕,〔現物・先物〕,〔売買・動き〕の各項目から 1 つずつ選んで作った 3 つ組であり，各文に対して少なくとも 1 つのトピックラベルが人手で付与されている．各項目の選択肢を表 1 に示す．

例えば「年末年始の米株式相場の下落を受けて投資家のリスク回避姿勢がやや強まり、売りが優勢だった。」には “米全体/現物/動き, 投資家/その他/その他, 日経平均/現物/売買” が付与されている．

付与されたトピック情報を考慮した生成を行うため、デコーダの LSTM の各時点にトピック情報を与える．具体的には、各時点の入力として直前の単語の埋め込みベクトル $\boldsymbol{v}$ とトピックラベルを埋め込んだベクトル $\boldsymbol{I}^k$ を結合したベクトルを用いる．つまり，第 k 文目の時点 t におけるデコーダの隠れ層の状態 $\boldsymbol{s}_t^k$ は，トピック埋め込みベクトル $\boldsymbol{I}^k$，直前の単語埋め込み $\boldsymbol{v}_{t-1}^k$，直前の隠れ層の状態 $\boldsymbol{s}_{t-1}^k$ を用いて次のように計算される:

$$\boldsymbol{s}_t^k = \mathrm{LSTM}\left(\left[\boldsymbol{I}^k; \boldsymbol{v}_{t-1}^k \right], \boldsymbol{s}_{t-1}^k \right). \tag{3}$$

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 $\boldsymbol{s}_t^k$ および重み $\boldsymbol{W}_s^k$ を用いて，$\mathrm{softmax}\left(\boldsymbol{W}_s^k \boldsymbol{s}_t^k\right)$ となる．

なお，上で述べたように，各文は複数のトピックを持ちうる．その場合，埋め込みベクトル $\boldsymbol{I}^k$ は，それぞれのトピックラベルの埋め込みベクトルを加算することで得られる．

3.3 数値表現を表す汎化タグの拡張

時系列株価データからテキスト中で言及された価格を適切に出力するために，テキスト中の数値表現を汎化タグに置換する．Aoki et al. (2018) が提案した汎化タグに対して，本研究で使用したデータに合うように国債

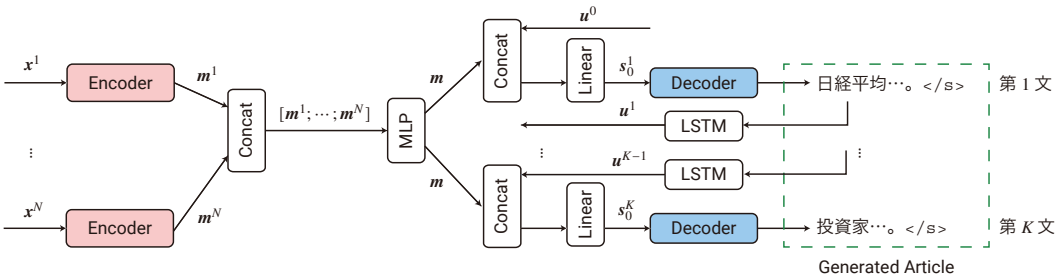

図 1: 提案モデルの全体概要図.

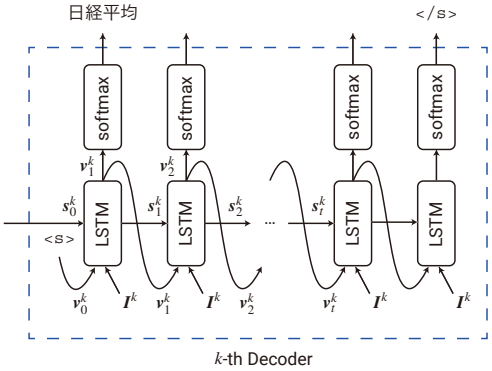

図 2: 第 k 文を生成するデコーダの概要図.

の利回りを表す<yield/>や四半期を表す<season/>を追加し，拡張を行った．

4 実験

4.1 データ

実験には時系列数値データとして，Thomson Reuters DataScope Select[*1]から 2012 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など 9 種類の時系列数値データを用いた[*2]．ま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ニュース記事から，一日の要約ニュースのみを抽出し，各文に 1 人のアノテータがトピックラベルを付与した．2012 年 1 月から 2015 年 9 月のデータを学習データ，2015 年 11 月から 2016 年 3 月のデータを開発データ，2016 年 4 月から 6 月のデータを評価データとして使用した．

4.2 実験設定

株価等の時系列数値データをベクトルへ変換する各エンコーダの隠れ状態の次元は，$\boldsymbol{x}_{\text{short}}$ に対しては 64，

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

[*2] 日経平均株価，東証株価指数，ダウ平均株価，上海総合指数，USD/JPY，USD/EUR，日本国債 2 年物，日本国債 10 年物，米国債 10 年物

表 2: 評価データにおける BLEU (%)

		追加情報		
		なし	トピック	出現箇所
注意機構	なし	11.9	13.4	9.9
	あり	–	14.9	13.7

$\boldsymbol{x}_{\text{long}}$ に対しては 128，前文のテキストデータに対しては 128 とした．これらの隠れ状態を結合した後，層数 2，隠れ状態の次元が 256 の MLP（多層パーセプトロン）を用いた．デコーダの隠れ状態の次元は 256 であり，トピック情報埋め込みベクトルの次元は 64 とした．またエンコーダ・デコーダにおける単語埋め込みベクトルの次元は 128 である．各パラメータの最適化手法には Adam (Kingma and Ba, 2015) を使用し，学習率は 5×10^{-4} とした．ミニバッチのサイズは 100，エポック数は 80 とした．

学習時に，開発データで計算された BLEU が連続して下がった場合には学習を終了する．そして，各エポックのモデルの中で，開発データに対する 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．

4.3 評価方法

評価実験として，トピックラベル情報を用いない設定，すなわち各 $\boldsymbol{I}^k$ を $\mathbf{0}$ とした場合，またトピックラベル情報の代わりに文の出現箇所を追加情報として用いる，すなわち各 $\boldsymbol{I}^k$ を $[0, \ldots, \overbrace{1}^{k}, \ldots, 0]$ で置き換えた場合の比較を行った．また，隠れ状態 $\boldsymbol{m}$ に対して，attention メカニズムによる注意機構を用いた設定も評価実験として行った．評価指標として BLEU を用い，モデルによって生成された各テキストと，実際の概況テキストの各文の一致度を測った．

4.4 実験結果

実験では，時系列数値データの表現手法および，トピックラベルなど追加情報の有用性について評価を行

表 3: 生成された文の例．記事中の出現箇所は 2 文目である．参照文に付与されたトピックラベルは，“米全体/現物/動き, その他/その他/その他, 日経平均/現物/動き, 投資家/その他/その他, 日経平均/現物/売買” である．

追加情報	生成文
(参照文)	前日の米株高やニューヨーク原油先物相場での下げ止まりを受け、反発して始まったが、大引けにかけて期末の持ち高調整の売りが出て相場を押し下げた。
追加情報なし	前日の米株式相場が下落した流れを引き継いで、朝方は買いが先行した。
文の出現箇所	前日の米株式相場が上昇したことを受けて、朝方から買いが先行した。
トピックラベル (参照文に付与されたもの)	前日の米株高を受けて投資家心理が改善し、日本株も売りが先行した。
トピックラベル (“米全体/現物/動き”)	前日の米株式相場が上昇した流れを引き継いだ。
トピックラベル (“円ドル/現物/動き”)	前日の海外市場で円安・ドル高が進んだ流れを引き継いだ。
トピックラベル (“投資家/その他/その他, 円ドル/現物/動き”)	前日の海外市場で米景気の減速懸念が強まり、円売り・ドル買いが進んだ流れを引き継いだ。

う．実験結果を表 2 および表 3 に示す．

表 2 によると，追加情報としてトピック情報を用いることで，BLEU の値が大きく向上することが分かった．また，追加情報を用いず生成されたテキストは正解文と比較して情報量が少なく，一様な文を生成する傾向があるが，追加情報を用いて生成されたテキスト，特にトピック情報を用いて生成されたテキストは多様で，トピック情報を考慮したテキストを生成できた．さらに表 3 の生成結果のように，過去のデータにないトピックラベルを入力として与えた場合もトピック情報を考慮することが出来た．

しかし，生成された文のうち，市況データの動向とテキスト表現が適切ではなく，数値データは「上昇」しているのに「下落」というテキスト表現を用いてしまうといった結果も見られたため，特定の表現において損失関数を重みづけするといった学習方法の提案が必要であると考えられる．

5 結論

本研究では，複数の市況データの概要を示すテキスト生成において，トピック情報を追加情報として用いることで，生成テキストの内容を制御するとともに，インタラクティブにテキストを生成する手法を提案した．評価実験では，トピック情報を追加することで，生成テキストの精度が向上することが示された．今後の課題として，生成されたテキストの人手評価および学習時に生成テキストが数値データの動向と一致していない場合に損失が大きくなるように，人工的に生成した負例を用いた学習方法の適用などが挙げられる．

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) の委託業務の結果得られたものです．

参考文献

Aoki, Tatsuya, Akira Miyazawa, Tatsuya Ishigaki, Keiichi Goshima, Kasumi Aoki, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao (2018) “Generating Market Comments Referring to External Resources,” in *Proc. of INLG 2018*, pp. 135–139.

Kingma, Diederik P. and Jimmy Ba (2015) “Adam: A Method for Stochastic Optimization,” in *Proc. of ICLR 2015*.

Murakami, Soichiro, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao (2017) “Learning to Generate Market Comments from Stock Prices,” in *Proc. of ACL 2017*, pp. 1374–1384.

Peng, Nanyun, Marjan Ghazvininejad, Jonathan May, and Kevin Knight (2018) “Towards Controllable Story Generation,” in *Proc. of First Workshop on Storytelling 2018*, pp. 43–49: Association for Computational Linguistics.

Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley (2010) “Automatic Keyword Extraction from Individual Documents,” in *Text Mining. Applications and Theory*, pp. 1–20.

Yao, Lili, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan (2018) “Plan-And-Write: Towards Better Automatic Storytelling,” *arXiv: 1811.05701v2*.

漆原理乃・小林一郎 (2018) 「人の動作および物体認識に基づく動画像からの文生成」,『言語処理学第 24 回会年次大会 (2018)』, 160–163 頁.

村上総一郎・笹野遼平・高村大也・奥村学 (2017) 「数値予報マップからの天気予報コメントの自動生成」,『言語処理学会第 23 回年次大会 (2017)』, 1121–1124 頁.