

市況コメント生成のための少数事例選択

川原田 将之[†] 石垣 達也[†] 高村 大也[†][†] 産業技術総合研究所

masayuki.kawarada.m@gmail.com

{ishigaki.tatsuya,takamura.hiroya}@aist.go.jp

概要

株価の数値時系列データからその値動きを説明する市況コメントを生成するタスクにおいて、プロンプトに含める少数ショット事例の選択方法が性能向上のために重要であることを述べる。従来、少数ショット事例は無作為に選択されており、その選択方法に関しては十分に研究されてこなかった。本研究では、類似する数値時系列データからのコメントや同一時点でのコメントの傾向が類似しているという特性を活用し、これらを少数ショット事例としてプロンプトに含める手法を提案する。評価実験において、ランダムに選択された事例を使用するベースラインと比較し、提案手法が代表的な評価指標である BLEU, METEOR, BERTScore をいずれも向上させた。

1 はじめに

市況コメント生成は、日経平均の値動きを表現した数値時系列データを入力とし、値動きを説明するテキストを出力する Data-to-text タスクの一つである。このタスクでは、従来、事前学習済みのニューラルネットワークを学習データを用いて、ファインチューニングする手法が主流であった [1]。近年は GPT 等の大規模事前学習言語モデルにプロンプトを与え言語生成する少数ショット学習 (few-shot learning) が注目されている [2]。そこで、本研究では、市況コメント生成を少数ショット学習で解く問題設定に着目し、プロンプトに含めるべき少数ショット事例の選択することで性能を向上させる手法を提案する。

少数ショット学習は、文書分類、質問応答、機械翻訳など幅広い言語処理タスクに適用され活発に研究されている。これらの研究では、「プロンプトの記述形式」、「プロンプトに含める少数ショット事例の選び方」、「プロンプト内での少数ショット事例の

提示順」が性能に大きく影響することが指摘されている [3]。市況コメント生成についての既存研究では、プロンプトの記述形式に着目するものが存在する [2] が、少数ショット事例に着目した研究は存在しない。既存の少数ショット事例の選択手法では、極性分析や文書分類において文の埋め込み表現の類似度を用いる手法、機械翻訳において高速な類似文字列探索手法を用いることで学習データセットから類似事例を探索するアイデアが提案されている [3, 4]。一方、市況コメント生成では入力が数値時系列データであるため、既存手法をそのまま適用することができない。

そこで、本研究では、図 1 に示すように、市況コメント生成タスクにおける 2 つの特徴に着目した類似事例探索を行い、プロンプトに含める少数ショット事例の選択を行う手法を提案する。1 つ目は、入力が数値時系列データである特徴に着目し、時系列同士の相関係数に基づき事例を選択する手法である。2 つ目は、入力の数値時系列データが時刻と対応しているという特徴に着目し、配信時刻の近さに基づいて事例を選択する手法である。

既存データを用いた実験において、2 つの手法は共に、従来のファインチューニングによる手法および、無作為に少数ショット事例を選択しプロンプトに含める手法よりも優れた性能を示した。

2 手法

タスク定義: 本研究では、短期系列と長期系列という 2 つの数値時系列データを入力として市況コメントを生成するタスクを扱う。1 日の日経平均株価の値動きを示した短期系列 S_N 、過去の日経平均株価の終値を示した系列を長期系列 L_M とする。短期系列には、現時刻から時間を遡って、等間隔に記録された N 個のデータが含まれている。短期系列 S_N は株価が記録された時刻を表す系列 T_N と対応しており、それぞれ、 $S_N = \{s_1, \dots, s_N\}$, $T_N = \{t_1, \dots, t_N\}$



図1 提案手法の処理手順. 1) テスト事例の時刻情報や値動きを手がかりに学習データから類似事例を探索し, 2) 少数ショット事例としてプロンプトの中で用いる.

と表される. ここで, t_1 , s_1 は, 現在の時刻と株価をそれぞれ表している. また, 長期系列において, M 日前までの系列が含まれるとすると, 長期系列は $L_M = \{l_1, \dots, l_M\}$ と表される. 市況コメント生成は S_N と L_M を入力として, 現在時刻 t_1 における市況コメントを出力するタスクである.

プロンプト: 図2に, 少数ショット学習に用いるプロンプトを示す. まず, 1行目に「現在時刻における市況コメント」を生成するようタスクについて説明する指示を記述する. 「###」に続く2行目以降に, 少数ショット事例とコメント生成対象とする数値時系列データを記述する. 短期系列のプロンプト表現では, まず冒頭に「時刻 日経平均株価」という見出しを与える. その後, 左カラムが時刻を表し, 右カラムが対応する株価となる表形式のプロンプトとして記述する. 長期系列については, 「日 日経平均株価終値」という見出し行に続き, 左カラムに「1日前」, 「2日前」等の生成対象時刻から遡る日数を記述し, 右カラムに終値を記述する. 図2において, 少数ショット事例は, 予測対象データと区別し T_N^{shot} と S_N^{shot} のように, 添え字の「shot」を付加して記載する. 「Output」以降の部分には, ショットの正解コメントである y^{shot} を記載する. 少数ショット事例についてプロンプトを記述した後, 生成対象とする短期系列, 長期系列を同様に表形式で記述し, 最後に「Output:」と記述し, 言語モデルに続きを生成させる.

少数事例の選択: 一般に, 少数ショット学習では, 学習データから $(T_N^{\text{shot}}, S_N^{\text{shot}}, L_M^{\text{shot}}, y^{\text{shot}})$ の組を無作為に抽出して, 少数ショット事例としてプロンプトに含める. 本研究では, 数値時系列データが持つ特性を活用して, 同時刻の事例と類似系列の事例の2つの方法によって少数事例の選択を行う:

同時刻の事例 この方法では, 予測対象データのコメントが発出された時刻 t_1 の同時刻の事例を学習データの中から無作為に取得する. 同時刻に発出される市況コメントは, 類似している場合が多く, ショットに含めることで品質の良い市況コメントが生成できることが期待できる.

類似系列の事例 この方法では, 予測対象データの短期系列 S_N と類似する系列を学習データから取得し, 少数ショット事例としてプロンプトに含める. これは, 類似する数値時系列データから生成される市況コメントは, 類似しているという仮説に基づいている. 本研究では, 系列の類似度を測る指標として, ピアソンの相関係数を用いる. 相関係数が高かった系列を順に並び替え, ショット数に応じて上位からプロンプトに含める.

3 実験

本節では, 本研究で行った実験について, データセット, 比較手法, 評価方法の順に記載する.

3.1 データセット

本研究では, 既存データ [1, 5, 6] を使用した. このデータセットには, 数値時系列データと市況コメントが対になった 18,489 事例が含まれている. 数値時系列データは IBI-Square ¹⁾ から 2010 年 12 月から 2016 年 9 月のデータを取得したものであり, 市況コメントは Nikkei Quick News から取得したものである ²⁾.

学習データ, 検証データ, テストデータがそれぞれ

1) <http://www.ibi-square.jp/index.html>
 2) データは契約を結べばダウンロード可能である. データの再現性を確保するため, 前処理のためのソースコードを公開している.

```
現在の時刻における市況コメントを<comment>市況コメント
</comment>の形式で出力してください。
```

```
###
```

```
Input:
```

```
時刻 日経平均株価
```

```
 $t_1^{shot}$   $s_1^{shot}$ 
```

```
⋮
```

```
 $t_N^{shot}$   $s_N^{shot}$ 
```

```
日 日経平均株価終値
```

```
1 日前  $t_1^{shot}$ 
```

```
⋮
```

```
 $M$  日前  $t_M^{shot}$ 
```

```
Output:  $y^{shot}$ 
```

```
###
```

```
Input:
```

```
時刻 日経平均株価
```

```
 $t_1$   $s_1$ 
```

```
⋮
```

```
 $t_N$   $s_N$ 
```

```
日 日経平均株価終値
```

```
1 日前  $t_1$ 
```

```
⋮
```

```
 $M$  日前  $t_M$ 
```

```
Output:
```

図2 少数ショット学習に用いるプロンプトの例

れ、15,035 事例、1,759 事例、1,695 事例となるように分割した。本研究では、既存研究に従い、短期系列を1日の取引の中で5分毎に記録された日経225の株価 ($N = 62$) とし、長期系列を過去7日間の株価の終値 ($M = 7$) とした。

3.2 比較手法

ファインチューニングによる手法 (BART)

Murakami ら [1] によるモデルをベースとしつつ、言語モデルを用いて拡張したモデルで実験を行った。彼らモデルでは、数値時系列データを多層パーセプトロン (MLP) に入力しベクトルに変換した後、そのベクトルを LSTM の Encoder-Decoder に入力することでテキストを出力している。我々のモデルでは、近年の Data-to-Text タスクの研究でよく用いられる事前学習済みの Encoder-Decoder モデルである BART [7]³⁾ を用いた。BART には、事前学習された単語埋め込み層が存在するが、本研究で扱うのは数値時系列データであるため、これらを用いずに、未学習の多層パーセプトロン (MLP) を新たに用意した⁴⁾。MLP を用いて短期系列と長期系列の2つのデータを768の固定サイズのベクトルに変換

3) <https://huggingface.co/stockmark/bart-base-japanese-news>

4) 単語埋め込み層以外のエンコーダ層とデコーダ層は事前学習済みのパラメータを用いた。

し、⁵⁾ BART を用いたテキスト生成を行う。

ゼロショット学習と少数ショット学習

LLM として、対話型大規模言語モデルである GPT-3.5 を使用した⁶⁾。少数ショット学習では、図2に示したプロンプトの形でそのまま使用するが、ゼロショット学習では、少数事例の記述を省略し、予測対象データのみを記述する。予備実験において、ショットに使用されるデータによってモデル性能が大きく変動することが判明したため、異なる seed 値を用いて事例を10回抽出して実験し、それらのスコアの平均を報告する。

3.3 評価手法

先行研究に従い、自動評価指標として、BLEU, METEOR, BERTScore を用いる。なお、BERTScore の算出には F1 スコアを用いる。BLEU は多くの研究で利用されているが、表層的な単語のみを考慮するため十分とは言えない。意味的な類似性の観点においても評価するため、BERT の埋め込み表現を利用する BERTScore も用いて評価を行う。

4 結果

表1に結果を示す。表の上段に LLM を用いた少数ショット学習での値を示す。少数ショット学習での結果については、ショット数を0, 3, 5, 10と増やした場合の性能を報告する(0がゼロショットに対応する)。表の下段は、ファインチューニングによる手法の性能である。

少数ショット事例を含めることの効果について述べる。市況コメント生成タスクはゼロショットでは BLEU, METEOR, BERTScore の値がそれぞれ 0.01, 0.48, 60.30 と非常に低い値を示し、十分な性能が得られないことがわかる。3ショットにすることで、無作為抽出するベースラインの値はそれぞれ 8.04, 24.85, 73.14 に大きく上昇する。よって、少数ショット事例をプロンプトに含めることが必要であり、少数ショット事例の選択手法についての研究は重要である。

同時刻の事例を含める提案手法の効果について述べる。学習データ中の同時刻に発出されたコメントをプロンプトに含めると、3ショットの設定において BLEU, METEOR, BERTScore の値がそれぞ

5) BART の埋め込み層のサイズが768であるため、このサイズを使用した。

6) 具体的には、OpenAI の API である gpt-3.5-turbo-16k を使用している。

れ, 13.11, 30.34, 74.86 に向上した. 無作為に選択しプロンプトを作成するベースラインでの値は, それぞれ 8.04, 24.85, 73.14 であるから, この提案手法によりすべての評価指標での大幅な性能向上を確認した. 少数ショット事例を 10 ショットに増やすと, 各指標で 17.46, 36.33, 76.40 にまで向上した.

学習データに含まれる類似系列を探索する提案手法も同様に無作為に抽出するベースラインより良い性能を示した. 特に 10 ショットを含める設定で, 類似系列を探索する手法は 17.66, 36.52, 76.42 とすべての評価指標でもっとも高い値を示した.

従来, 事実上の標準であったファインチューニングによる手法 (BART) との比較について述べる. BART は, 3 つの指標において 11.41, 30.90, 75.94 と無作為に 10 ショットの少数ショット事例を抽出する手法 (10.67, 28.40, 74.49) よりも高く, 少数ショット事例による手法開発は市況コメント生成に対しては難しい問題であることが分かる. 一方, 提案手法である「同時刻の事例」と「類似系列の事例」を選択する提案手法では, BART よりも良い性能を示した. この結果は, 従来大規模な学習データを用いたファインチューニングする手法が事実上の標準であった Data-to-text タスクにおいて, 時刻の活用といったドメイン知識や類似系列の探索と組み合わせたプロンプトの工夫により, 良好な結果が得られることを示すため興味深い.

5 関連研究

市況コメント生成は, Data-to-text の設定の一つである. 従来, 表 [8, 9], グラフ [10, 11], タブルの集合 [12], 数値時系列データ [1, 13, 14, 15] 等の非言語データを対象に研究されている. 本研究で扱う数値時系列データは他の設定と異なり, プロンプトとして記述するための表現方法が明らかではない. 市況コメント生成は従来, 大規模な学習データを用いてニューラルネットワークをファインチューニングする手法が一般的だが, ゼロショット学習や少数ショット学習についての研究も始まっている [2].

少数ショット学習における事例選択手法については, 極性分析, 表からの説明テキスト生成, 質問応答, 機械翻訳などの問題において, 入力テキストの埋め込み表現を用いて学習データ中の類似事例を探索する手法が提案されている [3]. 過去の類似事例をプロンプトに含めることで性能が向上する報告は, 我々が本研究で得た知見と一致する. 一方, こ

少数ショット学習による手法			
	BLEU	METEOR	BERTScore
0 ショット			
含める事例なし	0.01	0.48	60.30
3 ショット			
無作為抽出	8.04	24.85	73.14
同時刻の事例	13.11	30.34	74.86
類似系列の事例	13.14	31.27	75.07
5 ショット			
無作為抽出	9.49	26.69	73.83
同時刻の事例	13.17	31.30	75.08
類似系列の事例	13.17	30.33	74.86
10 ショット			
無作為抽出	10.68	28.40	74.49
同時刻の事例	17.46	36.33	76.40
類似系列の事例	17.66	36.52	76.42
ファインチューニングによる手法			
BART	11.41	30.90	75.94

表 1 各比較手法の BLEU, METEOR, BERTScore. 提案手法である「同時刻の事例」は無作為に抽出するベースライン手法よりも良い性能を示した., 「値動きが類似する事例」を選択する手法は同時刻の事例を選ぶ手法よりも良い性能を示した.

これらの既存研究では, RoBERTa などの事前学習済みのトークン埋め込みを用いるため, 市況コメント生成のような数値時系列データにそのまま適用できない. 機械翻訳向けには開発セットでの性能が高くなるように事例を選ぶ手法が提案されている [4]. この手法は市況コメント生成にも適用できる可能性があり, 今後適用可能性を検討したい.

6 結論

本研究では, 大規模言語モデルを用いた少数ショット学習による市況コメント生成タスクに着目し, 少数ショット事例を選択する手法を提案した. 実験より, 同時刻に記述された過去のコメント, 過去の類似する値動きにおいて記述されたコメントをプロンプトに含めることで, 良好な言語生成能力を得られることを確かめた. ショット事例の選択を適切に行った少数ショット学習による提案手法が, 従来, 事実上の標準で良い性能を示していたファインチューニングによる手法よりも良い性能を示すことが分かった.

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) の助成事業 (JPNP20006) による支援の結果得られたものである。

参考文献

- [1] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. Learning to generate market comments from stock prices. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1374–1384, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [2] 川原田将之, 石垣達也, 高村大也. 少数ショット学習による時系列数値データからの市況コメント生成. 情報処理学会第 258 回自然言語処理研究会, pp. 1–8. 情報処理学会, 2023 年 11 月.
- [3] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, **Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures**, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.
- [4] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 8857–8873, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Tatsuya Aoki, Akira Miyazawa, Tatsuya Ishigaki, Keiichi Goshima, Kasumi Aoki, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. Generating market comments referring to external resources. In **Proceedings of the 11th International Conference on Natural Language Generation**, pp. 135–139, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics.
- [6] Yumi Hamazono, Tatsuya Ishigaki, Yusuke Miyao, Hiroya Takamura, and Ichiro Kobayashi. Unpredictable attributes in market comment generation. In **Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation**, pp. 217–226, Shanghai, China, 11 2021. Association for Computational Linguistics.
- [7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.
- [8] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. In **Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence**, AAAI’19/IAAI’19/EAAI’19. AAAI Press, 2019.
- [9] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1203–1213, Austin, Texas, November 2016. Association for Computational Linguistics.
- [10] Xuefeng Bai, Yulong Chen, and Yue Zhang. Graph pre-training for AMR parsing and generation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6001–6015, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [11] Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. Neural AMR: Sequence-to-sequence models for parsing and generation. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 146–157, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [12] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. The WebNLG challenge: Generating text from RDF data. In **Proceedings of the 10th International Conference on Natural Language Generation**, pp. 124–133, Santiago de Compostela, Spain, September 2017. Association for Computational Linguistics.
- [13] Ernie Chang, Alisa Kovtunova, Stefan Borgwardt, Vera Demberg, Kathryn Chapman, and Hui-Syuan Yeh. Logic-guided message generation from raw real-time sensor data. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6899–6908, Marseille, France, June 2022. European Language Resources Association.
- [14] Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Generating racing game commentary from vision, language, and structured data. In **Proceedings of the 14th International Conference on Natural Language Generation**, pp. 103–113, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics.
- [15] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4005–4023, Dublin, Ireland, May 2022. Association for Computational Linguistics.