小規模データセットを用いた景気の基調判断文自動生成

木村柚里 [1*]　末廣徹 [1,2]　稲垣真太郎 [1]

[1] みずほ証券金融市場調査部

[2] 法政大学大学院経済学研究科経済学専攻

{yuri.kimura, toru.suehiro, shintaro.inagaki}@mizuho-sc.com

1. はじめに

近年、質問応答やテーブル検索など、さまざまな領域で構造化データを活用することに焦点を当てた多数の研究が行われている．構造化データの活用をテーマとした研究分野の 1 つに、データを自然言語に変換し、理解可能性を拡張する“Data-to-Text” [1]が挙げられる．この領域における試みは、スポーツ実況[2]から医療診断[3]、金融[4,5]など、幅広い分野において応用可能性が検討されている．Data-to-text に関する研究においては、再帰的ニューラルネットワークまたは Attention 機構を用いた encoder-decoder モデル [6, 7]を応用した手法が多い．しかしこれらの研究は主に大規模データの使用が前提となっている．Data-to-Text に限らず深層学習に係る領域では、大規模データを扱ったモデルの構築が前提となっており、1000 以下の小規模なサイズのデータセットに対する応用はほとんど見られない．

本稿では、日本の景気変動に一致すると考えられている複数の経済指標を用いて、日本政府の景気の基調判断文を生成することができるかどうかの検証を行った．本稿が分析対象とした月次の経済指標と毎月の日本政府の基調判断文は先行研究で用いられてきたニュース記事や市場データを用いた研究と比較してデータセットが小規模である．このような論点を扱う場合、大量のデータセットを用意することは困難である．しかし、先行研究では大規模なデータが必須であることが暗黙の了解となっており、小規模なデータへの応用に関しては未だ発展途上である．本稿では、深層学習モデルについて、事前学習やデータ拡張を行わずに精度を向上させるアプローチを検証しており、小規模データセットに対する応用可能性という面からも、当該分野における貢献があると考えられる．

また、金融経済の数値データから文章を生成するという試み自体は過去にも分析された例があるものの、過去の研究は市場参加者に対する情報提供や業務効率化という側面が強い．一方、本稿が分析対象とする日本政府の基調判断文の「クセ」を学習することで、事前に日本政府の景気認識の変化やそれに伴う政策変更の予想に役立てることができるなど、フォワードルッキングな利用も大きなモチベーションとなっている．

2. 関連研究

小規模データセットへの深層学習の応用に関する議論において近年取り上げられているのは転移学習、Fine-tuning といった大規模データセットによる事前学習モデルの応用やデータ拡張の領域が多い．裏を返せば上記アプローチの他に小規模データセットへの応用が取り上げられる機会は、現状かなり限られている．しかし、事前学習やデータ拡張を使わずとも小規模なデータセットに深層学習モデルを適用するためのアプローチとして有力な手掛かりを提示する研究も複数存在する．

例えば Olson らの研究[8]は分類タスクにおいて深層学習モデルは小規模なデータセットにおいてもランダムフォレストと同等程度の精度を出すことを示している．Olson らは、この理由を深層学習モデルが小規模なネットワークの集合であると捉えることで、大規模ニューラルネットがランダムフォレストと同じように複数のモデルのアンサンブルとして予測を行っているためであると考えている．彼らの実験では UCI Machine Learning Repository から小規模データの 116 データセットを用いてニューラルネットとランダムフォレストの精度比較も行なわれたが、大体のデータセットで深層学習モデルはランダムフォレストに近い分類精度を示した．また Salman らによる研究[9]では 1000 の MNIST データにおいて、分類の難しいサンプルを取り除くことで、分類予測モ

* 連絡先: みずほ証券株式会社金融市場調査部
〒100-0004 東京都千代田区大手町 1-5-1 大手町ファーストスクエア
E-mail: yuri.kimura@mizuho-sc.com

デルの精度を向上させた例を紹介している．彼らのアプローチは特にサイズが限られているデータに対して有効であり、さらに統計的手法に依らない予測モデルに対して大きな効果があると述べている．

すなわち、小規模なデータセットに対しても深層学習は適応する可能性が十分にある．そしてより精度の高いモデルを構築するためには、学習においてターゲットの特徴をつかみやすいような良質な入力データが必要であることを示唆する．つまり、処理タスクに関わる領域に明るい人による事前知識を活用した訓練データの構築がモデルの精度を大きく押し上げる可能性が高い．

本稿では encoder-decoder モデルによる構造化データからの文書生成タスクが小規模データセットに対して、事前学習モデルやデータ拡張を使用せず、事前知識によるデータの前処理のみよってどの程度向上するか検証した．

3. 使用データ

内閣府月例経済報告の 1998 年 1 月から 2019 年 8 月で示される景気の基調判断のテキストデータを用いた．月例経済報告は景気に対する日本政府の公式見解を示す資料であり、「月例経済報告等に関する関係閣僚会議」において経済財政政策担当大臣を中心に議論されて内容が決まる．表 1 に「景気の基調判断」の例を示す．「景気の基調判断」では景気動向指数も参考にされるものの、CI 一致指数などによって機械的に基調判断が決定するわけではない．従って、日本政府の基調判断はどのような基準で作成されているのか不透明な点が多いことで知られている．

基調判断の文章の構成についても決まりがあるわけではなく、「景気は、回復している」などシンプルな構成の場合もあれば、「景気は、輸出や生産に弱さがみられるが、持ち直している」など様々な付帯条件が記されることもある．とはいえ、ほとんどのケースにおいて基調判断は「景気は」という語で始まる文章であり、文の最後は「弱含んでいる」「改善に足踏みがみられる」「持ち直しの動きがみられる」などのように結論が示されている．このように、一定のパターンをもったデータセットであることから、データの数に制約がある下でも一定の精度で文章の生成が可能であると考えられる．また、分析に用いたデータセットでは、「東日本大震災の影響により」や「復興需要等を背景として、」といった経済データだけでは判断しにくいとみられる文言はセンテンスから取り除いた．他にも、輸出や個人消費、生産と言った細かい表現については景気全体の水準や方向性とは直接的な関係がないことから、これもデータから除いた．さらに、2 センテンス以上から構成されているコメントについては、初めのセンテンスのみを取り出してターゲットデータとした。表 2 に例を示す．

文章生成の元となる経済データは、景気動向指数の CI 一致指数の個別系列（9 系列）を用いた．CI（Composite Index）は生産、消費、雇用など経済活動での重要かつ景気に敏感に反応する様々な指標を統合して１つの指標にすることによって、景気の現状把握を行うために作成されている．

表１：景気の基調判断例

景気は、おおむね横ばいとなっているが、イラク情勢等から不透明感が増している。
景気は、おおむね横ばいとなっている。株価やアメリカ経済の動向など、我が国の景気を巡る環境に変化の兆しがみられる。
景気は、依然として厳しい状況にあるものの、復興需要等を背景として、緩やかに回復しつつある。

表２：ターゲットデータ例

景気は、おおむね横ばいとなっているが、不透明感が増している
景気は、おおむね横ばいとなっている
景気は、依然として厳しい状況にあるものの、緩やかに回復しつつある

景気動向指数には、景気に対して先行して動く先行指数、ほぼ一致して動く一致指数、遅れて動く遅行指数の３つの指数があるが、内閣府経済社会総合研究所では、一致指数の各採用系列から作られるヒストリカル DI 等に基づき、景気動向指数研究会での議論を踏まえ、景気循環の転換点である景気基準日付（景気の山・谷）を設定している．日本政府の景気に対する基調判断と、景気動向指数研究会の判断は必ずしも一致しないものの、一般的に景気の基調判断を決めるデータセットとしては過不足がないものと言える．

指標データに関して、本稿では当月を除く過去 12 ヶ月分を入力データとする．当月分については、月例経済報告が公表された段階で経済指標が発表されていないものが多いことから、基調判断を決める際の参考にすることはできないはずである．データセットの合計は 260 で、ここから 80%を訓練データ、20%を検証用データとして学習を行う．すなわちモデル入力データのサイズは 208 である．これは Olson らの実験[8]で使用された 116 種のデータケースのう

表３:景気動向指数 CI 一致指数例

年	C1(100)	C2(100)	C3(100)	C4 (100)	C5(100)	C6 (%)	C7(%)	C8(億円)	C9(倍)
2016/4	99.3	98.9	98.9	98.5	98.3	-0.9	-5.3	143,035	1.33
2016/5	98.5	98.2	99.5	97.9	97.2	-2.1	-6.7	142,132	1.35
2016/6	99.2	99.3	98.3	98.7	97.8	-1.3	-7.3	141,230	1.36

C1:生産指数(鉱工業), C2:鉱工業用生産財出荷指数, C3:耐久消費財出荷指数, C4:所定外労働時間指数(調査産業計), C5:投資財出荷指数(除輸送機械), C6:商業販売(小売業、前年同月比), C7:商業販売額(卸売業、前年同月比), C8:営業利益(全産業), C9:有効求人倍率(除学卒)

ち 25 パーセンタイルに相当する程度の大きさであり、検証にあたって極端に小さいとは断言すべきではなく、検証の価値は十分にあるといえる.

4. 実験

本稿では文書生成タスクについて encoder-decoder モデルを使用する．Encoder、Decoder 内部のアーキテクチャにはどちらも GRU を使用する. GRU とはゲート機構を持つ回帰型ニューラルネットワークであり、LSTM と同等の精度が期待されるが、出力ゲートが特に小規模なデータセットに対しては GRU が有効とされている[10].
本稿の検証では以下 3 つのモデルの比較を行う：

1. 指標データに前処理を施さないモデル (model1)
2. 指標データに前処理を施したモデル(model2)
3. 指標データに前処理を施し、0.2 の Dropout を適用したもの(model3)

BLEU-4 スコアを偏りなく比較するため、テキストデータのトークナイズは同じ手順でおこなう．訓練データ、検証データ、テストデータは各モデルについて同じものを使用する.

4.1 テキストデータの前処理

助詞、助動詞、非自立的動詞、接続詞は区切らず、手前で出現したトークンに接続する．下にトークンリストの例を示す：

［〝景気は〞, 〝、〞, 〝緩やかに〞, 〝回復しつつ〞, 〝ある〞］

作成したトークンをリストに格納し、ID 化することで埋め込み処理を可能にした.

4.2 指標データの前処理

入力データである指標データに関する前処理として、以下のように月次差分の標準化処理を行う：

$$\mathrm{m}_{ij} = x_{ij} - x_{(\mathrm{i}-1)\mathrm{j}}$$

$$\mathrm{m}'_{ij} = \frac{m_{ij} - \overline{m_j}}{\sigma_j}$$

ここで、i は時系列方向の添え字であり、jはそれぞれの指標を示す添え字である．$\overline{m_j}$ は指標 j における平均値でありσ_{j} は指標 j における標準偏差である. すなわち、ここで行っている処理は各データの前月との差分の標準化である．ここで得た $\mathrm{m}'_{\mathrm{ij}}$ を x_{ij} に代わるエンコーダへの入力データとして使用する.

4.3 実験結果

実験における条件は、model1、model2、model3 のそれぞれについて共通化した．具体的には、パラメータの最適化手法には Adam を使用し、ミニバッチのサイズを 12、encoder の隠れ状態の次元を 256、epoch 数を 100 とした．形態素解析器には Janome[1]、モデルの実装には Pytorch[2]を使用した.

モデルの評価には BLEU-4 スコアを用いた．データサイズが限られているため評価データを用いない評価を試みた．model1, model2, model3 について、validation data に対する BLEU-4 スコアの最大値を取得するという試行をそれぞれ 10 回行い、その平均値を求めた．結果を示した表 4 によると、model1 と比較して model2 と model3 のスコアが大幅に上昇していることがわかった．指標データの前処理によってモデルの精度が飛躍的に向上していると推測される. また図 1 ではモデルの各 epoch に対する 10 回の試行の平均値を図示した．model1 に比べ model2、model3 は epoch を重ねるごとに順調に精度が向上していることが見受けられる．一方 dropout による正則化に

[1] https://mocobeta.github.io/janome/

[2] https://pytorch.org/

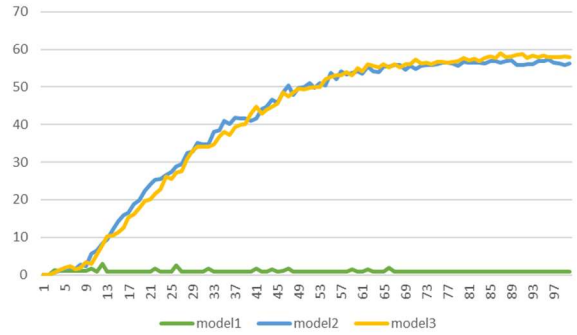
ついて、model2 と model3 を比較したところ、特段差異が無いことが観察された．

以上の結果は Olson ら[8]や Salman ら[9]の研究結果と一致すると考えられる．すなわち、深層学習を用いた encoder-decoder モデルにおいて、数百程度の小規模データセットのみを用いた学習であっても適切な入力データを用いることで充足的に高い精度のモデルを作ることが不可能ではないということを示した．Olson ら[8]や Salman ら[9]による議論は分類予測モデルに留まったが、今回はこれが encoder-decoder モデルによるテキスト生成タスクにおいて当てはまる可能性を示す例となったと考えられる．

表 4　各モデル BLEU-4 スコア

	model1	model2	model3
BLEU-4	3.1	59.2	60.3

図 1 :BLEU-4 スコア比較



各モデルの validation データにおける BLEU-4 スコアの推移．
縦軸：BLEU-4 スコア、横軸：epoch 数

5. 結論

本稿では、日本の経済指標と日本政府の基調判断のテキストデータを用いて、小規模なサイズのデータセットに対する深層学習の応用可能性について検証した．その結果、小規模なデータセットであっても一定の前処理を行うことで、十分な精度で深層学習を用いた Data-to-Text のタスクを行うことが可能であることが分かった．

具体的には、内閣府の景気動向指数の CI 一致指数に用いられている 9 つの月次の経済指標を用いて、内閣府月例経済報告で示される日本政府の景気の基調判断のテキストデータを生成するタスクの検証を、GRU を用いた encoder-decoder モデルによって行った．前処理の有効性を検証するため、指標データに前処理を施さないモデル(model1)、指標データに前処理を施したモデル(model2)、指標データに前処理を施し、dropout を適用したもの(model3)、の 3 つを比較した結果、model1 と比べ model2 や model3 では十分な精度でテキストデータの生成が可能であることが分かった．ただし model2 と model3 のおいては dropout を用いた正則化による精度の向上は極めて限定的であるという結果を得た．

本稿の結果は実用的なモデルの作成において、必ずしも大規模なデータセットを用意しなければならないという制約から解放されうる可能性があることを提示した．今回提示したものはあくまで一例に過ぎないが、今後多様なデータセットに対して深層学習モデルを活用しうる一助となるだろう．むろん、前処理の有用性についてはデータセットによる面も大きいため、どのようなデータセットで応用可能かについての検証が今後の課題となる．また、今回は一定の前処理によって精度の向上する可能性を示したが、ここで示したアプローチに対して文書拡張や大規模言語モデルの fine-tuning など事前学習モデルを応用した技術との対比になども今後議論が必要である．

参考文献

[1] Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65-170.

[2] Iso, H., Uehara, Y., Ishigaki, T., Noji, H., Aramaki, E., Kobayashi, I., ... & Takamura, H. (2019). Learning to Select, Track, and Generate for Data-to-Text. *arXiv preprint arXiv:1907.09699*.

[3] Pauws, S., Gatt, A., Krahmer, E., Reiter, E.: Making Effective Use of Healthcare Data Using Data-to-Text Technology: Methodologies and Applications, pp. 119– 145 (01 2019)

[4] Murakami, S., Watanabe, A., Miyazawa, A., Goshima, K., Yanase, T., Takamura, H., & Miyao, Y. (2017). 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)* (Vol. 1, pp. 1374-1384).

[5] Aoki, T., Miyazawa, A., Ishigaki, T., Goshima, K., Aoki, K., Kobayashi, I., ... & Miyao, Y. (2018, November). Generating market comments referring to external resources. In *Proceedings of the 11th International Conference on Natural Language Generation* (pp. 135-139).

[6] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in NIPS*.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

[8] Olson, M., Wyner, A., & Berk, R. (2018). Modern neural networks generalize on small data sets. In Advances in Neural Information Processing Systems (pp. 3619-3628).

[9] Salman, S., & Liu, X. (2019). Overfitting mechanism and avoidance in deep neural networks. arXiv preprint arXiv:1901.06566.

[10] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.355*