

二つの時系列データの関係を記述する自然言語文生成

理学専攻・情報科学コース
2340671
中野 由加子

1 はじめに

Data-to-Text は、表や時系列チャートなど多様な表現形式での数値データの重要な部分について文生成を行う研究を指す。大量のデータから重要な部分を抽出して自然言語で説明をすることは人々のデータへの理解の助けとなるため Data-to-Text の研究は注目されている。多くの Data-to-text の研究は、数値データが観測されたドメインに特化した自然言語記述と同様な自然言語文を生成することを対象にしている [1]。一方で、時系列データそのものが表現する意味や複数の時系列データ間の関係を捉えることも同じように重要にもかかわらず、そのような研究は少数であり、かつ一つの時系列データを対象とする [2]。同じ時間軸の複数の時系列データが与えられた場合、各データの個別の挙動に加え、これらのデータ間に生じる相互関係が分析対象として浮上する。相互関係の把握は、複数データ間での運動性や影響力を評価するための重要な要素となる。本論文では、二つの時系列データの関係を捉えるため Transformer のクロス注意機構を拡張したモデルを用いて二つの時系列データの関係を説明する文を生成する手法を提案する。研究 1 では、二つの時系列データの動向についての説明文生成、研究 2 においては二つの時系列データの挙動についての質問応答文生成を行う。研究 3 では、人工データで学習した研究 2 のモデルの実測データへの転移学習の有効性について検証を行う。

2 データセット

研究 1,2 では説明対象として設定をした挙動をとる時系列データを作成し、多様な分野に応用できるように正解文で特定の分野の専門用語の使用は避けることで、ドメインフリーの人工データセットを構築した。

2.1 時系列データにおける説明対象

時系列データにおいて人が顕著に感じる動向として、「増加」、「減少」、「頂上」、「凹み」の 4 つが挙げられる [2]。本研究では、二つの時系列データの説明対象の動向関係を上記の数値的变化量がある 4 種類の動向の組み合わせで構成する⁰(表 1)。時系列データ長を 60 とし、0-19 を「初期」、20-39 を「中期」、40-59 を「末期」とする。これらに加え、研究 2,3 では二つの時系列データの動向の数値的变化量の比較結果を“more”, “less”, “considerable”, “slight” という単語を用いて説明をする。また、時系列データが数値的变化量を取る動向を取り始める地点と取り終わる地点の比較結果を“before”, “after”等の単語を用いて説明をする。

2.2 データセット概要

データセットの質問文(研究 2,3 のみで使用)と正解文の作成にはテンプレートを用いており、質問文と時系列データの挙動を踏まえて動向などの単語(図 1,2 の例において色がついている単語)を置き換えて作成する。研究 1 用のデータセット 1 と 2、研究 2 用のデー

表 1: 動向の関係

協調	増加/増加, 減少/減少, 頂上/頂上, 凹み/凹み
反対	増加/減少, 減少/増加, 頂上/凹み, 凹み/頂上
片方平坦	増加/平坦, 減少/平坦, 頂上/平坦, 凹み/平坦
	平坦/増加, 平坦/減少, 平坦/頂上, 平坦/凹み

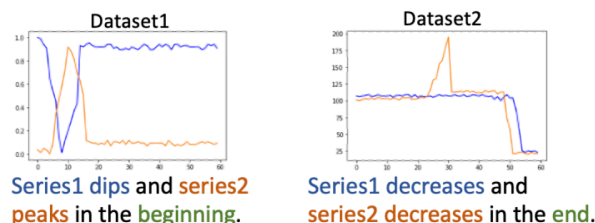


図 1: データセット 1 と 2 の例

タセット 3-5、研究 3 用の実測データを用いたデータセットについての説明を以下に示す。

研究 1 データセット 1 では図 1 の左のように説明対象の動向関係が 3 つの時期のうちの 1 つにおいて取られる時系列データを、データセット 2 では右のように説明対象の動向に加えて説明対象でない「片方平坦」の挙動をとる時系列データを構築する。

研究 2 精度向上のために提案モデルの学習を 3 段階に分けて行い、それぞれの訓練のためのデータセット(図 2)を作成し、両者に同じ番号を振った。データセット 3 は、時系列データ、特定の時期について問う質問文、正解回答文で構成する。データセット 4 は、時系列データと全ての挙動について説明をする正解文で構成する。データセット 5 は、時系列データ、挙動についての質問文、正解文で構成する。全てのデータセットにおける時系列データは、初期、中期、末期それぞれにおいて表 1 のうちの一つの挙動をとるようにした。

研究 3 実測データでの評価のために、コロナ新規感染者数のデータ、気候が異なる世界の 5 つの都市の気温データ、株価の終値のデータを使用する。それぞれのデータごとに約 30 ずつランダムに抽出して、正解文を割り当てた。

3 提案モデル

提案モデルのアーキテクチャは、入力データの関係を捉えるために Transformer [3] のエンコーダ部分を拡張した。Encoder1 はそれぞれの時系列データ自身を分析し、Encoder2 は二つの時系列データの関係性を検証する。エンコーダへの入力時系列データは、データの動向を表すためにエンベディングベクトルに変換され、query, key, value ベクトルが transformer の self-attention メカニズムにより作られる。左の Encoder1 において、全ての 3 つのベクトルは series1 で右の Encoder1 においては series2 となる。一方で、Encoder2 においては key, value は Encoder1 の入力ベクトルと同じで query は隣の Encoder2 と交換される。研究 1 では、質問文を用いないため Encoder3 は使用せずに Encoder2 の出力がデコーダへの入力となる。研究 2,3

⁰片方平坦は研究 2,3 でのみ説明対象とする

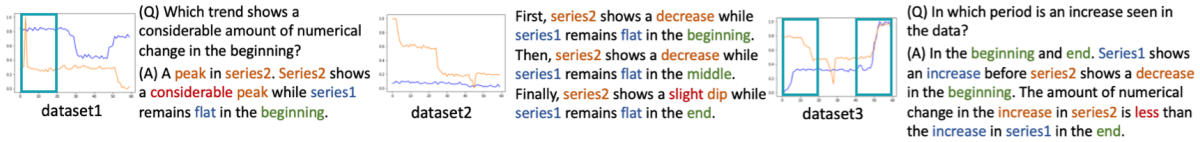


図 2: データセット 3-5 の例

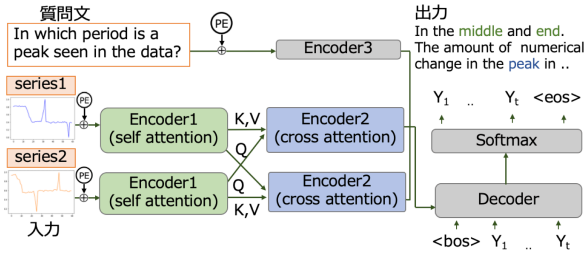


図 3: 提案モデルの概要

では質問文を入力とするため、Encoder3は質問文を解析し、データのどこに注目すべきかの洞察を得る。デコーダは通常の Transformer デコーダと同じとする。

4 実験

4.1 訓練手法

研究 1 実験 1 では 4 つの実験を行ない、それぞれの実験を (訓練データセット番号)-(テストデータセット番号) を用いて表す。実験 1-1,2-2 では、データセットから学習した動向について正しく文生成を行うことができるかを評価する。実験 1-2 では、実験 1-1 で訓練をしたモデルを用いて観測対象でない動向も含むデータセット 2 について文生成を行うことでモデルの頑健性を見る。実験 2-2 では実験 2-1 で訓練をしたモデルで、説明対象としていない動向を含んだデータにおける共通部分などから捉えたい動向を学習し正しく文生成ができているかを評価する。

研究 2 訓練 3 は質問文の反映を目的とし、時期を指定した質問文への回答を通してデータの着目すべき視点を誘導して質問文を反映した文生成の手法を学習する。訓練 4 は単語ベクトルの獲得を目的とし、データにおける全ての説明対象の挙動についての説明文生成を通して、データにおける挙動とそれを意味する単語の対応関係を学習する。訓練 5 は質問回答文生成を目的とし上記二つの訓練を踏まえて、データの挙動についての質問への回答文生成の手法を学習する。

4.2 評価手法

研究 1 では生成文の質を評価するための自動評価指標として BLEU と perplexity (ppl.) と、単語の正解率として生成文と正解文における 2 つの動向 (T), 2 つのうちどちらかの動向 (1T), 時期 (P), 2 つの動向と時期 (T&P) の一致率を用いて評価を行う。研究 2,3 では BLEU, ROUGE-L, BertScore-Precision, METEOR と、単語の正解率として、2 つの動向, 時期, 程度 (D), 時間的關係 (C) とこれらの組み合わせの一致率を評価対象とする。研究 2,3 における評価は値が大きいほど精度が良いことを意味する。

4.3 結果

研究 1 実験 1-1, 2-1 の高い精度から、多くの場合で人工データから捉えたい動向と時期を学習して正しい文生成を行えたことが確認できる (表 2)。実験 1-2 については、実験 1-1 に比べて大きく精度が下がっている。

表 2: 研究 1 結果

Ex	BLEU	T&P	P	T	1T	ppl.
1-1	0.998	0.995	0.995	1.0	1.0	10.9
2-1	0.982	0.971	0.977	0.987	0.988	4.04
1-2	0.677	0.509	0.707	0.752	0.787	3.67
2-2	0.988	0.978	0.978	0.994	0.994	4.57

表 3: 実測データでの評価結果

データ	BLEU	ROUGE	METEOR	BERT	T&P&D&C	T	P	D	C
気温	0.536	0.756	0.684	0.949	0.551	0.533	0.625	0.558	0.286
株価	0.619	0.746	0.722	0.955	0.654	0.674	0.761	0.542	0.607
コロナ	0.594	0.832	0.713	0.959	0.631	0.634	0.698	0.566	0.531
全て	0.586	0.767	0.714	0.954	0.613	0.613	0.700	0.545	0.508

時期や動向の単語が観測対象でない動向のものになってしまっている場合が多く見られた。一方で実験 2-2 では実験 2-1 よりも全体的に精度が高くなっている。観測対象としていない動向を含んだデータにおける共通部分などから二つの時系列データにおける動向を正しく学習して文生成を行うことができたことが確認できる。

研究 2 訓練 3&4&5 は常に訓練 5 よりも外側にあることから、3 段階の訓練を行うことで精度が上がることを確認される²。生成文が不正解であるもののうち、動向や程度などを正しく説明できていない例も見られたが、回答対象の動向を一部見逃している場合や質問が反映できていない例も多く見られた。

研究 3 研究 2 の結果から正解率は下がっているが (表 3), 人工データを通して学習した挙動に近い場合は正しく回答文生成を行うことが確認できた。

5 まとめ

本研究は、二つの時系列データの関係を記述する自然言語文生成を行うために、ドメインフリーの人工データセットを用いてクロス注意機構を拡張したモデルの訓練を行なった。研究 1,2 から、多くの場合でドメインフリーの人工データセットを通して学習した挙動を正しく捉え、説明文生成が可能であることが確認された。研究 3 の正解率は十分に高いものではなかったが、人工データで学習をした挙動に近いデータの場合は正しく回答ができることが確認された。文生成を行う分野のデータの挙動に寄せた人工データセットの使用や実測データを用いた fine-tuning を行うことで、より高い精度が期待される。今後はより多様な表現での説明文生成を行なっていきたい。

参考文献

- [1] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model, 2020.
- [2] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Truth-conditional captions for time series data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 719–733, 2021.
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

²実験におけるテストデータは最後の訓練番号となり、例えば訓練 3&4 であれば訓練データがデータセット 3 と 4 で、テストデータはデータセット 4 となる。

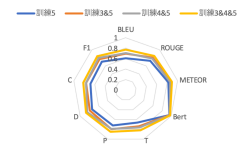


図 4: 研究 2 の結果