二つの時系列データの関係を記述する自然言語文生成への取り組み

中野 由加子　(指導教員：小林 一郎)

1　はじめに

近年，様々な分野における数値データの収集が容易になり，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について自然言語で記述する Data-to-Text の研究が注目されている．多くの先行研究は，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．そのため，数値データ以外にもその数値が観測されたイベントを表す表題や発話者などのデータ発生の情報源といったコンテクスト情報を追加した自然言語文生成が提案されている [1]．一方で，分析した数値解析結果を可視化ではなく言語化することによって，そのデータの内容を容易に把握することが可能になり，また対話システムや分析システムに組み込むことも可能になると考えられる．本研究では，二つの時系列データの関係性を説明する文生成を対象とする．

2　データセット構築

二つの時系列データの関係性を説明するために，二つの時系列データの動向が共に increase，decrease，peak，dip である場合と，一方が increase で他方が decrease，一方が peak で他方が dip である場合の動向を学習させ，動向と時期を正しく捉えた文が生成できているかについて評価を行う．データセットは，二つの時系列データ，関係性を示すカテゴリ，正解文のペアとし，時系列データの種類が異なる 2 種類のデータセットを用意した．データセット 1 は，捉えたい動向のみが含まれており，データセット 2 は，捉えたい動向に加え 4 種類のうちの 1 つの動向がノイズとして含まれている．

時系列データ　データセット 1 では，対象とする時系列データの関係となる 8 種類の動向とそれが生起する 3 種類の時期の 24 通りのデータを作成した (Appendix 参照)．全体の時間幅を 60 と設定し，「beginning」の場合は，0-19，「middle」の場合は 20-39，「end」の場合は 40-59 の範囲に二つの時系列データの関係を示す動向が収まるようにしてデータを構築した．選ばれた時期以外では，時系列データにほとんど変動がないようにした．動向が「increase」・「decrease」，時期が「beginning」である場合は，0-19 の範囲内で increase・decrease の動向がある時系列データを作成する (図 1 参照)．データセット 2 では，片方の時系列データに，捉えたい動向がある時期以外の時期のどちらかに任意の動向を加え，24 種類の時系列データのペアに対して 16 通りのノイズを加えた (Appendix 参照)．図 2 は，捉えたい動向がともに「dip」，時期が「middle」で series1 (青色の時系列データ) に「increase」の動向が「beginning」に追加された例となっている．作成したデータは，二つの時系列データの値が $[0, 1]$ に収まるように正規化を行ってからモデルに入力する．

関係カテゴリ　エンコーダとデコーダを分けて訓練する際に用いる．二つの時系列データの関係性を捉えられるようにエンコーダを訓練をするために，カテゴリ

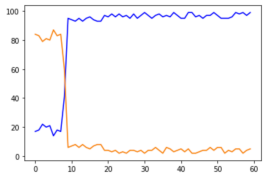


図 1: データセット 1

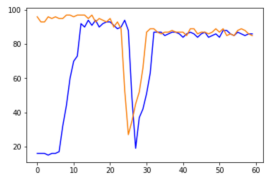


図 2: データセット 2

を用意した．上述した，動向 8 種類，時期 3 種類の組み合わせ 24 種類を 0-23 の数字で表したものをカテゴリとして用いる．

正解文　動向と時期について言及する文を生成文の正解としている．データセットの作成を簡単にするために，“Series1 (trend1) and series2 (trend2) in the (period).” というテンプレートを使用し，動向 (trend) 4 種類と時期 (period) 3 種類についての単語を置き換えることで 24 種類の正解文を生成した．

3　提案モデル

3.1　概要

提案するモデルの概要を図 3 に示す．

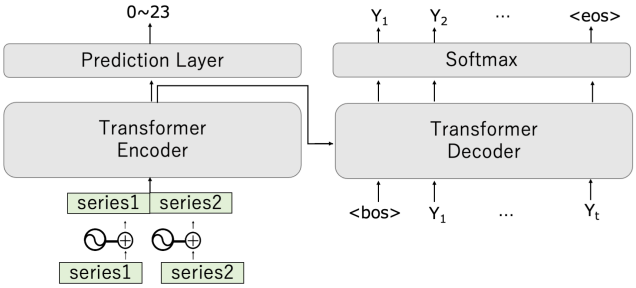
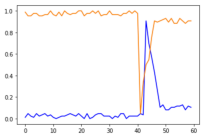
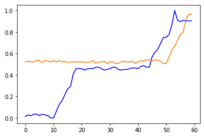


図 3: 時系列データの関係説明文生成モデルの概要

モデルは Transformer [4] をベースに構築されている．入力は二つの時系列データであり，Transformer を用いた多次元時系列データの異常検知を行う手法である TranAD [3] を参考に形式を決めている（3.2 節にて後述）．また，2 つの動向の比較を可能にするために positional encoding をそれぞれに追加することにより時刻を共通化している．本論文では，モデルの訓練にエンコーダとデコーダを別々で訓練する手法 (手法 A) と End-to-End で訓練する手法（手法 B）の 2 つを用いた．前者については，時系列データの特徴を上手くエンコーディングするために，先行研究 [1] のエンコーダの訓練手法を参考に，エンコーダからの出力を受け，上述した時系列データの動向と時期によって表される 24 種類のカテゴリを識別する層（図 3 中の Prediction Layer）を導入し，カテゴリの識別を訓練することにより，時系列データの関係を正しく捉えた埋め込み表現をデコーダに渡している．デコーダは Transformer のデコーダを採用している．エンコーダおよびデコーダの訓練時のハイパーパラメータの設定を表 1 に示す．



生成結果
Series1 peaks and series2 dips in the end.(peak・dip・end)



生成結果:
Series1 increases and series2 increases in the end.(increase・increase・end)

図 4: データセット 1,2 の生成文と (生成カテゴリ) の正解例

表 1: 訓練設定

バッチサイズ	8
Embedding	128 次元
隠れ層	512 次元
損失関数	cross entropy
勾配法	Adam
学習率	0.0001
ドロップアウト	0.1
ウィンドウサイズ	8

表 2: 実験 1,2,3 結果

実験	手法	BLEU	ppl.	2 動向	1 動向	時期	2 動向/時期	カテゴリ
1	A	98.0	4.07	0.999	0.999	0.973	0.972	0.993
1	B	99.3	3.75	0.999	0.999	0.964	0.963	-
2	A	95.3	4.02	0.979	0.981	0.921	0.914	0.885
2	B	93.7	3.75	0.962	0.966	0.910	0.890	-
3	A	97.2	4.17	0.999	0.999	0.938	0.938	0.949
3	B	98.5	4.35	0.999	0.999	0.967	0.967	-

3.2 時系列データの入力形式

時系列データの振る舞いを捉えるために,「スライディングウィンドウ（sliding window）」と呼ばれる, ある程度の範囲を時間方向に少しずつシフトさせた特徴量によって時系列データを表現することが多い. 本研究においてもそのことを踏襲し, 入力形式として時系列データにスライディングウィンドウを適用したものを採用する. ウィンドウサイズを K とし, 時刻 t における値を $W_t = \{x_{t-K+1}, x_{t-K+2}, \ldots, x_t\}$ とする. このとき, 入力は, w_t のリストとなり, 長さ T の時系列データの場合, $W = \{W_1, W_2, \ldots, W_T\}$ と表される. この形式をとることで, 変化量がわかりやすくなり, 時系列データの動向が捉えやすくなる. 本研究では, この値に続けて, それぞれの値の差を追加することで, モデルに与える情報を増やし, 精度を高めた.

4 実験

4.1 実験設定

データセット 1,2 ともに, データを 96,000 個作成し, その内, 85%を訓練用データ, 5%を評価用データ, 10%をテスト用データとした. 10,000step 訓練を行なった内, 1,000step ごとに評価データを用いて文生成を行い, BLEU 値が最大の時のモデルを実験に使用した. 実験 1,2 では, それぞれデータセット 1,2 の訓練用データで, 捉えたい二つの時系列データの動向をモデルに学習させ, データセット 1,2 のテスト用データで評価を行う. 実験 3 では, データセット 2 の訓練用データで訓練したモデルでデータセット 1 のテスト用データについて文生成を行うことで, 4 つの動向のうちの 1 つ (以下「ノイズ」と呼ぶ) が加えられた 16 種類のデータから共通部分を捉えることで, 捉えたい動向について文生成をできているかを評価する.

4.2 評価方法

生成文の質を評価するために, BLEU [2] と perplexity (ppl.) を用いた. 動向・時期を捉えられているかを確認するために, 生成した文と正解文において, 2 つの動向, 2 つのうちどちらかの動向, 時期, 2 つの動向と時期を表す単語が一致しているかどうかを判定した. これらに加えて, 手法 A については Prediction Layer の判定結果も評価対象とする.

4.3 実験結果及び考察

全体として, 手法 1 ・ 2 の精度にあまり差は見られなかった (表 2 参照).

カテゴリ判定と生成文の正解率について 生成文において動向と時期の正解率は, カテゴリ判定の正解率と比べてあまり差は見られなかった. 多くの時系列データについて, エンコーダで正しく捉えられていることが確認できた.

実験 1,2 BLEU 値が非常に高くなっているが, 今回はテンプレートを用いて正解文を生成しているおり文同士に大差がなく, また動向と時期を高い精度で捉えられているためだと考えられる. 実験 1 に比べて実験 2 の精度が全体的に下がっており, ノイズの影響を受けたと考えられる. 生成文における動向の単語が, ノイズの単語になっているものが見られた.

実験 3 実験 2 と比べて全体的に精度が上がっている. ノイズが含まれているデータから捉えたい動向を学習したため, ノイズの含まれていないデータにおいて正しい文生成が可能となったと考えられる.

5 まとめ

本論文では, 二つの時系列データの関係を記述する自然言語文生成のために, 動向と時期について説明する文生成を行なった. 今後は, 動向と時期をより正確に捉えられるようモデルの精度向上を目指す.

参考文献

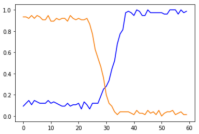
[1] Obeid, J. and Hoque, E.: Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model (2020).

[2] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 311–318, USA (2002), Association for Computational Linguistics.

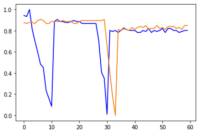
[3] Tuli, S., Casale, G. and Jennings, N. R.: TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data, *Proc. VLDB Endow.*, Vol. 15, No. 6, p. 1201–1214 (2022).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, in Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. eds., *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc. (2017).

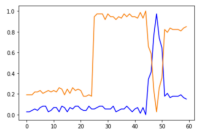
A　生成文正解例/正解カテゴリ



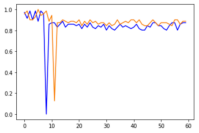
生成文: Series1 increases and series2 decreases in the middle. (increase・decrease・middle)



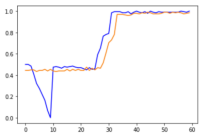
生成文: Series1 dips and series2 dips in the middle. (dip・dip・middle)



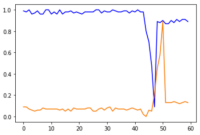
生成文: Series1 peaks and series2 dips in the end. (peak・dip・end)



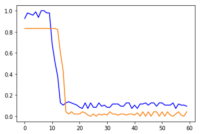
生成文: Series1 dips and series2 dips in the beginning. (dip・dip・beginning)



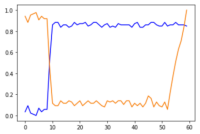
生成文: Series1 increases and series2 increases in the middle. (increase・increase・middle)



生成文: Series1 dips and series2 peaks in the end. (dip・peak・end)

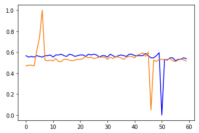


生成文: Series1 decreases and series2 decreases in the beginning. (decrease・decrease・beginning)

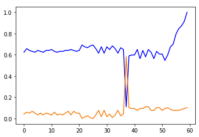


生成文: Series1 increases and series2 decreases in the beginning. (increase・decrease・beginning)

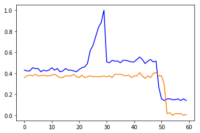
B　生成文不正解例/正解文例



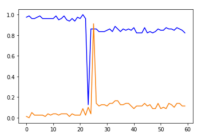
生成文: Series1 peaks and series2 peaks in the end.
正解文: Series1 dips and series2 dips in the end.



生成文: Series1 dips and series2 peaks in the beginning.
正解文: Series1 dips and series2 peaks in the middle.

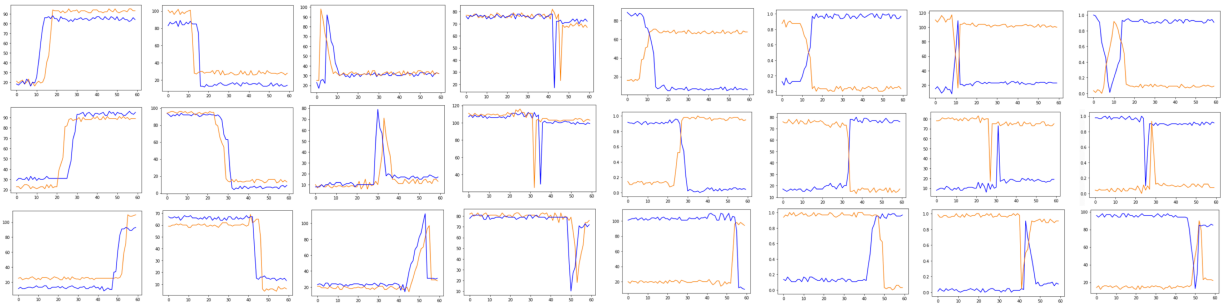


生成文: Series1 peaks and series2 peaks in the end.
正解文: Series1 decreases and series2 decreases in the end.



生成文: Series1 dips and series2 peaks in the beginning.
正解文: Series1 dips and series2 peaks in the middle.

C　データセット１ 全 24 種類



D　データセット２ peak・dip・middle に動向を一つ加えた場合の例 全 16 種類