3R-08

時系列データのパターンを考慮した言語モデルに基づく自然言語生成

青木花純 †
† お茶の水女子大学理学部情報科学科

小林一郎 ‡
‡ お茶の水女子大学 基幹研究院自然科学系

1 はじめに

近年，センサ等から観測される時系列数値データを様々な用途で利用する場面が増えている. しかし，時系列データをそのまま表示するだけでは，数値データの概要を人が把握するのは困難であり，人の理解を助けるために，テキスト表現等を用いた動向概要を付与することが多く行われている. そのため, 時系列数値データから動向概要を示すテキスト等を自動生成する技術への関心が高まっている. また，自然言語処理の分野においても，視覚情報として観測されるデータを時系列数値データとして処理し，テキスト生成する研究が盛んになっている [1, 2, 3]. 本研究では, 日経平均株価を例に，時系列数値データの動向概要を示すテキストの自動生成に取り組む.

2 時系列データのテキスト生成

2.1 概要

本研究では，過去に観測された時系列数値データのパターンと動向概要を示した文章内容の対応関係を学習し，文章から構築された適切な言語資源を利用することによって，観測された数値データの概要を表現するテキストを生成する. 図 1 に研究の概要を示す.

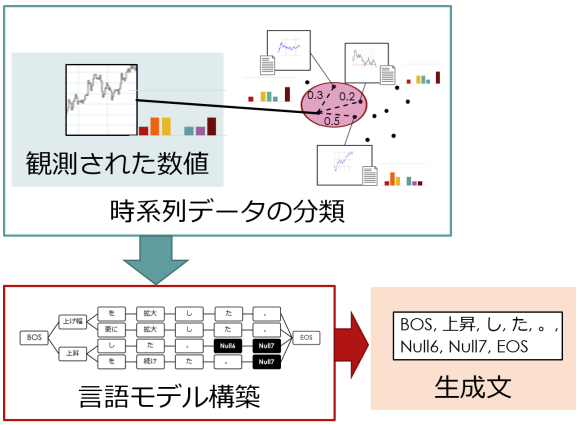


図 1: 研究概要図

Natural Language Generation based on Language Model considering Patterns of Time-series Data
†Kasumi AOKI(g1120501@is.ocha.ac.jp),
‡Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

まず，観測された時系列数値データと過去に観測された時系列数値データに対して Dynamic Time Warping 距離を類似度として用い，スペクトラルクラスタリング手法を適用し，任意の個数のクラスタに分類する. そして，観測された時系列数値データが属するクラスタに分類された各時系列数値データの動向内容を示した文書を言語資源とし，バイグラムモデルを作成する. その際，観測されたデータと類似度の高い時系列数値データの言語資源に重み付けを行う. このようにして生成した言語モデルを用いて，確率的に尤もらしい単語の組み合わせを決定し，観測された時系列数値データの動向概要を示すテキストを生成する.

2.2 時系列データの分類

まず, スペクトラルクラスタリングを用いた時系列データの分類において, 時系列同士の類似度には Dynamic Time Warping(DTW) 距離を用いた. 同じクラスタに分類された時系列データと対で収集した文書を言語モデルを構築する言語資源とする. 本研究では，類似度には Dynamic Time Warping(DTW) 距離を用いた.

2.3 言語モデルの構築と文生成

言語モデルとして，観測された時系列データと同クラスタの言語資源を用い, バイグラムモデルを構築し, 確率的に尤もらしいテキストを生成した. その際, クラスタ内の各時系列数値データと観測された時系列データとの類似度（DTW 距離）を基に各言語資源に重み付けを行う. テキスト生成には, 重み付けされて得られた言語資源からバイグラムを作成し, 動的計画法を用いて, 尤度が高くなる単語の組み合わせを得ることにより文を生成する. 尤度は長い文ほど低くなってしまうことから, 図 2 のように文長に左右されない言語モデルを構築するため，言語モデルを構築する各言語資源は仮想の単語 [null] を用いて文長を揃えた.

3 実験

本章では，上記に説明した手法を用いて，新たな日経平均株価の時系列数値データが与えられた際，その内容を説明するテキスト生成の実験について説明する.

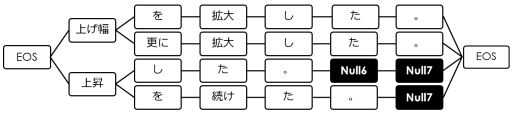


図 2: null を用いたバイグラムモデル

3.1 実験設定

時系列数値データは 4〜8 個のクラスタに分類されるとする. 株価の時系列数値データ，および言語モデルを構築する文章は前場，後場の各時間帯にわけて収集した. 実験に使用したテキストデータ *，および数値データ † は，2014 年 1 月 6 日〜2014 年 12 月 30 日に収集された 244 日分の 488 個のデータである．今回は収集したデータのうち，ランダムで選択したデータを新たに観測されたデータだとみなし，提案手法を適用した.

3.2 スペクトラルクラスタリング実行結果

提案手法を用いて，時系列数値データをスペクトラルクラスタリングした際の分類例を以下に示す.

表 1: スペクトラルクラスタリング結果

]

クラスタ数/ID	1	2	3	4	5	6	7	8
4	110	123	86	169	–	–	–	–
5	57	82	105	126	118	–	–	–
6	100	113	77	51	74	73	–	–
7	72	77	97	57	29	88	68	–
8	59	41	66	89	55	83	50	45

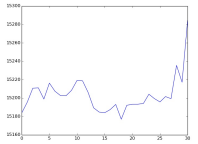
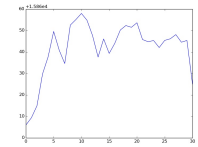
その後，クラスタリング結果および観測された時系列データとの類似度による重み付けによってバイグラムモデルを構築し，動的計画法を用いることで，株価数値データの概要を説明する尤もらしい文を生成した. 実行結果として, クラスタ数が 8 の場合に生成された文を, 時系列数値データおよび正解文とともに表 2 に示す.

3.3 考察

実験では，正解文と同様の文を生成できた事例もあったが，正解文と全く違う動向内容を説明するものも存在した. しかし，生成文と株価数値データを比較すると，生成文が全く違うとは言い切れない. これは正解文中に，前時間帯の情報，円相場との兼ね合い等の対応する時系列データの動向内容以外の情報が含まれているからだと考えられる. そのため，今後は生成文の評価をより沢山のデータを用いて行いたいと考えている．また，今回はクラスタリングのクラスタ数を 4〜8 と動的に設定し重み付けを行ったが, クラスタ数が少ないものほど使用する言語資源が多くなってしまうため，人手による評価も踏まえて, クラスタ数に応じた重み付けおよび, 最適なクラスタ数を決定したいと考えている.

*ADVFN:http://jp.advfn.com/より取得
†IBI-Square Stocks:http://www.ibi-square.jp/より取得

表 2: 言語モデルによる生成文

株価動向	生成文
	正解文:小動きとなった。
	• 小, 動き, と, なっ, た, 。, null7,···, null35, EOS • 上げ幅, を, 拡大, し, た, 。, null7,···, null35, EOS
	正解文: 一段高となった
	• 上げ幅, を, 拡大, し, た,。, null7, ···,null29, EOS • 一時，上げ幅, を, 拡大, し, た,。, null7, ···,null29, EOS

4 おわりに

本研究では，日経平均株価を対象に，観測された時系列データの概要を説明するテキストの自動生成に取り組んだ. 時系列数値データに対しクラスタリングを行い，必要な言語資源を決定することで構築した言語モデルから，動的計画法を用い，尤度の高い単語の組み合わせを得ることで文生成を行った. 数値データの動向内容の的確なテキスト表現を生成出来たものもあったが，該当時間帯の動向内容のみ表現しているため，正解文と異なるテキスト表現を生成してしまうものもあった. 統計的な評価尺度とともに，人手による評価も行っていきたいと考えている.

参考文献

[1] Gkatzia, D., Hastie, H. and Lemon, O. Finding middle ground Multi-objectuve Natural Language Generation from time-series data, the 14th European Association for Computational Linguistics, pp.210-214, 2014

[2] H., Banaee, M. U. Ahmed, A. Loutfi, A Framework for Automatic Text Generation of Trends in Physiological Time Series Data, IEEE Int. Conf. on Systems, Man, and Cybernetics, pp.3876-3881,2013.

[3] 小林瑞希, 小林一郎, 麻生英樹, 同画像中の人の動作を表現する確率的言語生成に関する取り組み, 第 27 回人工知能学会全国大会,2D5-OS-03b-3, 2013.

[4] Ulrike von Luxburg, Max Planck Institute for Biological Cybernetics Spr,spemannstr. 38, 72076 Tubinge, Germaniy “A Tutorial on Spectral Clustering”, Statics and Computing 17 (4), 2007.

[5] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis, A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts, In The University of Texas at Austin, Department of Computer Science. Technical Report TR-04-25, 2005.