時系列データの類似度に基づき重み付けされた言語モデルを用いた文生成

青木　花純　　小林　一郎
お茶の水女子大学理学部情報科学科　お茶の水女子大学 基幹研究院自然科学系
{g1120501,koba}@is.ocha.ac.jp

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，時系列データを表示する際には，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おり，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, 6]．本研究では，日経平均株価を例に，時系列データの類似度を基に重み付けされた言語モデルを生成し，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　時系列データからの文生成

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その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言語モデルを利用することによって，観測された時系列数値データの動向概要を表現するテキストを生成することを目的とする．図 1 に研究の概要を示す．

まず，新たに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 距離を時系列データ同士の類似度としてスペクトラルクラスタリングを適用し，任意の個数のクラスタに分類する．そして，新しく観測された時系列数値データと同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からバイグラムモデルを構築する．その際に，新しく観測された時系列数値データと同クラスタに分類された時系列データの類似度に応じて重み付けを行う．

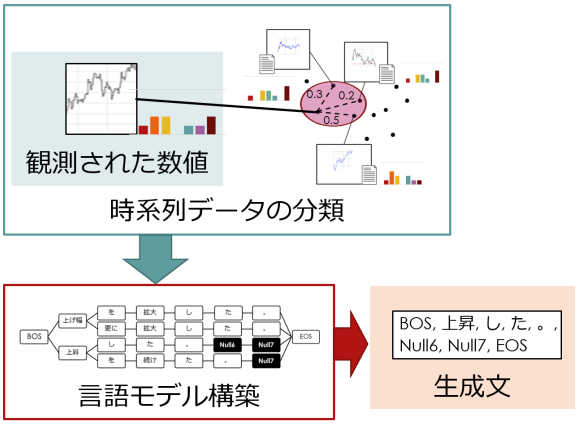


図 1: 研究概要図

以上のように生成したバイグラムモデルに対し，動的計画法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

時系列データの分類にはスペクトラルクラスタリングを用いた．スペクトラルクラスタリングは各データをノード，各データ間の類似度をノード間の距離として Normalized Cut を行う事で，各データをクラスタリングする手法である．本研究では，時系列データ同士の類似度には各時系列データの Dynamic Time Warping(DTW) 距離を用いた．DTW 距離とは時系列データの各点の距離を総当りで比較し，総計コストが最短となるパスでかかる総コストのことである． 観測された新しい時系列数値データと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．

2.3 観測データに対するバイグラムモデルの構築

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てバイグラムモデルを構築した. その際, 観測された時系列データと同クラスタ内の各時系列データの類似度（DTW 距離）を基に各言語資源に重み付けを行い，バイグラムモデルを構築した.

2.4 言語モデルによるテキスト生成

テキスト生成には，時系列データの類似度により重み付けされて得られたバイグラムモデルに対し，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．尤度は文長が長い文ほど低くなってしまうことから，文長に左右されないテキスト生成を実現するため，言語モデルを構築する際に，バイグラムモデルを構築する言語資源の最大文長に合わせて，各言語資源すべてに仮想の単語として番号付きの null ラベルを擬似単語として導入した.

図 2: 仮想単語 null の挿入

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を行い，評価を行う.

3.1 実験設定

今回使用する日経平均株価は動向内容が上昇，下落後安定などおおよそ 9 個に経験的に分類できると仮定し，実験ではその数の前後の数に分類されると想定し，時系列数値データは 6〜12 個にクラスタリングされるとした．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，後場の各時間帯に分けて収集した. 実験に使用したテキストデータ [1]，および数値データ [2] は，2014 年 1 月 6 日〜2014 年 12 月 30 日に収集された 244 日分の 488 個の前場，後場のデータである．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とみなし，提案手法を適用した.

[1]ADVFN:http://jp.advfn.com/より取得.
[2]IBI-Square Stocks:http://www.ibi-square.jp/より取得。

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表 1 に示す.

表 1: 時系列データの分類

クラスタ数/ID	1	2	3	4	5	6	7	8	9	10	11	12
6	100	113	77	51	74	73	–	–	–	–	–	–
7	72	77	97	57	29	88	68	–	–	–	–	–
8	59	41	66	89	55	83	50	45	–	–	–	–
9	63	56	51	52	50	57	55	52	52	–	–	–
10	58	30	59	49	66	48	46	44	40	48	–	–
11	33	43	49	35	74	31	37	61	43	37	45	–
12	37	42	37	49	49	38	40	24	26	57	52	37

その後，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.
実行結果の例として，クラスタ数が 4，8，12 の場合に重み付けを適用してバイグラムを作成し，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

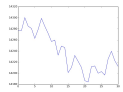
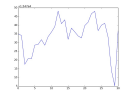
クラスタ数によらず，生成された文は比較的短い文が多かった．これは収集したテキストにおいて，短い文が比較的多いためであると考えられる．また，クラスタの数が少ないものほど，生成された文の精度が低いように感じた．今後は人主観に基づく評価とともに，正解データを用意し，BLEU などを用いて統計的な評価を踏まえて，生成文の精度評価をしていきたいと考えている.

4 おわりに

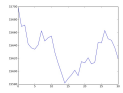
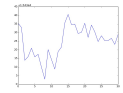
本研究では，日経平均株価を対象に，観測された時系列データの概要を説明するテキストの自動生成に取り組んだ．時系列数値データに対し DTW 距離に基づくクラスタリングを行い，新たに観測された時系列データと同クラスタに分類された時系列データのバイグラムに類似度を重み付けすることによりバイグラムモデルを構築し，そのバイグラムモデルに対して，動的計画法を用いることにより，尤度の高い単語の組み合わせを得ることで文生成を行った．時系列数値データの分類におけるクラスタ数や言語モデルを構築の際の重み付け方法を比較し，考察した．今後は時系列データのより正確な分類，バイグラムモデルの構築などを行い，精度の高い文生成を行いたいと考えている.

表 2: クラスタ数の変化に対する文生成結果

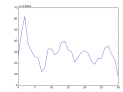
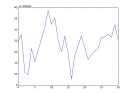
• クラスタ数:4

株価動向	生成文	尤度
 正解文:一時上げ幅を拡大。その後も概ね底堅い動きとなった。	• 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null27, EOS	1.68e-49
	• 下げ, 幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null27	6.55e-50
	• 一時, 下げ, 幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null26	9.21e-51
 正解文:小幅安水準で小動きとなった	• 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null29, EOS	2.73e-53
	• 一時, 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null29	7.33e-54
	• 一時, 下げ, 幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null28	9.39e-55

• クラスタ数:8

株価動向	生成文	尤度
 正解文:一時下げ幅を拡大したが、売り急ぐ動きは少なく、その後はやや下げ渋った。	• 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null31, EOS	8.23e-57
	• 一時, 下げ, 幅, を, 拡大, し, た, 。, null8, null9, null10, null11, ···, null31, EOS	8.23e-57
	• 一時, 下げ, 幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null31, EOS	7.03e-58
 正解文:小動きに終始した。	• 一時, 下げ幅, を, 拡大, し, た, 。, null9, null10, null11, null12, ···, null39, EOS	6.35e-75
	• 一時, 下げ幅, を, 拡大, し, た, 。, null8, null9, null10, null11, ···, null39	6.80e-76
	• 上げ, 幅, を, 拡大, し, た, が, 、, 一時, 上げ幅, を, 拡大, し, た, 。, null13, null14, null15, null16 ···, null38	1.14e-76

• クラスタ数:12

株価動向	生成文	尤度
 正解文:一時上げ幅を拡大したが、その後はやや伸び悩んだ水準で小動きとなった。	• 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null28, EOS	1.41e-50
	• 一時, 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null28, EOS	2.65e-51
	• 一時, 下げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null28, EOS	3.36e-52
 正解文:概ね底堅く推移した。	• 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null39, EOS	4.03e-72
	• 一時, 上げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null39, EOS	1.64e-72
	• 一時, 下げ幅, を, 拡大, し, た, 。, null7, null8, null9, null10, ···, null28, EOS	3.99e-73

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 (2013). 第 27 回人工知能学会全国大会,2D5-OS-03b-3, 2013.

[4] Ulrike von Luxburg "A Tutorial on Spectral Clustering" Max Planck Institute for Biological Cybernetics Spr,spemannstr. 38, 72076 Tubinge, Germaniy, 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

[6] 青木花純, 小林一郎, 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, 情報処理学会,2016