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　(指導教員：小林 一郎)

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る．このことから人の理解を助けるために，文章による動向概要を付与することが多く，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, 5]．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　時系列データを説明する文生成

2.1　概要

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，スペクトラルクラスタリングを適用し，特定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．その際使用する言語資源には，時系列データ同士の類似度に応じて重み付けを行う．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，テキストを生成する．

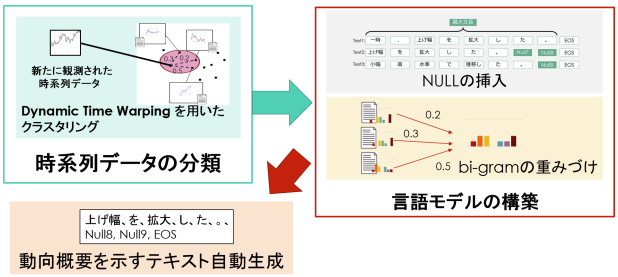


図 1: 研究概要図

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 [3] を用いた．スペクトラルクラスタリングは各データをグラフのノード，データ間の類似度をノード間の距離としてグラフ分割を行う事で，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時系列データ間の Dynamic Time Warping(DTW) 距離 [4] 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　バイグラムモデルの構築およびテキスト生成

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てバイグラムモデルを構築する．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，観測された時系列データを説明するためのバイグラムを生成する．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言語モデルを構築するために，モデルを構築する言語資源の最大文長に合わせて，各文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 [5]．

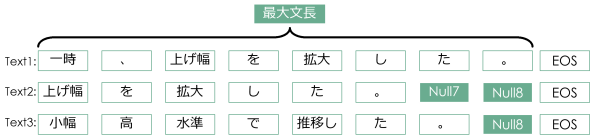


図 2: 仮想単語 null の挿入

3　実験

本章では，上記に説明した手法を用いて，新たな時系列数値データが与えられた際，動向概要を示すテキスト生成の実験を行い，評価を行う．

3.1　実験設定

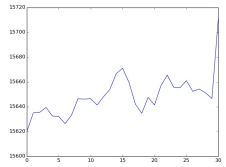
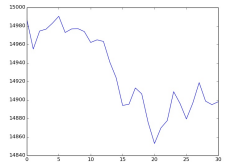
今回使用する日経平均株価は，動向内容が上昇，下落後安定など，おおよそ 9 個に経験的に分類できると仮定し，実験ではその数の前後の数を想定した 6〜12 個にクラスタリングされるとした．株価の時系列数値データ，および言語モデルを構築する文章は前場，後場の各時間帯に分けて収集した．実験に使用したテキストデータ [1]，および数値データ [2] は，2014 年 1 月 6 日〜2014 年 12 月 30 日に収集された 244 日分の 488 個の前場，後場のデータである．

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，新たに観測された時系列データと過去に観測された時系列データのクラスタリングを行い，言語モデル構築の際の言語資源の選択に利用した．クラスタ数によって選択された言語資源の数，および言語資源内の単語の種類，bi-gram の種類に違い

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

表 2: クラスタ数の変化に対する文生成結果

株価動向	クラスタ数	生成文	対数尤度
 【正解文】円相場を睨みながら、一進一退の相場となった。	6	上げ幅, を, 拡大, し, た, 。, null8, ···, null36, EOS	-152.01
	7	下げ, 幅, を, 拡大, し, た, 。, null9, ···, null27, EOS	-116.06
	8	上げ幅, を, 拡大, し, た, 。, null8, ···, null25, EOS	-110.41
	9	上げ幅, を, 拡大, し, た, 。, null8, ···, null29, EOS	-115.59
	10	上げ幅, を, 拡大, し, た, 。, null8, ···, null27, EOS	-120.22
	11	上げ幅, を, 拡大, し, た, 。, null8, ···, null29, EOS	-111.72
	12	上げ幅, を, 拡大, し, た, が, 、, 下げ, 幅, を, 拡大, し, た, 。, null16, ···, null25, EOS	-112.63
 【正解文】一時下げ幅を拡大した。	6	一時, 上げ 幅, を, 拡大, し, た, 。, null9, ···, null28, EOS	-133.90
	7	一時, 上げ 幅, を, 拡大, し, た, 。, null9, ···, null28, EOS	-132.81
	8	一時, 上げ幅, を, 拡大, し, た, 。, null9, ···, null28, EOS	-120.86
	9	下げ, 幅, を, 拡大, し, た, 。, null9, ···, null29, EOS	-117.75
	10	小, 動き, と, なっ, た, 。, null8, ···, null29, EOS	-106.76
	11	上げ幅, を, 拡大, し, た, 。, null8, ···, null27, EOS	-122.28
	12	一時, 下げ, 幅, を, 拡大, し, た, 。, null10, ···, null25, EOS	-123.14

が見られた．その結果を表 1 に示す．

表 1: 時系列データの分類

クラスタ数												
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

項目/クラスタ数	6	7	8	9	10	11	12
単語の種類	147	145	133	151	116	115	113
bi-gram の種類	344	334	285	315	230	237	230
DTW の平均値	0.66	0.62	0.68	0.61	0.70	0.60	0.58

3.3 生成テキストの評価

提案手法を用いて構築したバイグラムモデルに動的計画法を用いることで，観測された時系列データの動向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 6〜12 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．また，クラスタ数毎に生成された文の評価を表 3 に示す．

表 3: 生成文の評価

項目/クラスタ数	6	7	8	9	10	11	12
precision	0.53	0.54	0.52	0.52	0.52	0.52	0.53
recall	0.37	0.37	0.36	0.36	0.36	0.37	0.37
F1 value	0.41	0.42	0.41	0.40	0.40	0.41	0.42

3.4 考察

生成された文の中には正解文と全く違うような文を生成しているものもあったが，グラフの動向概要を示している文としては適切なことが多かった．これは，同じような時系列データでも動向概要の表現にばらつきがあるからだと考えられる．また，今回クラスタ数毎に生成された文の評価を行った．数値としては大きな違いは見られなかったものの，クラスタ数が多いものほど詳細に動向概要を示しているように思われた．

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，選択された言語資源に重み付けすることによりバイグラムモデルを構築し，動的計画法を用いることにより，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を比較し，評価指標を用いて生成文を評価し，結果を考察した．今後はクラスタ数や重み付けの調節や評価方法の見直しなどを行い，精度の高い文生成を行いたいと考えている．

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 EACL, 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] 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

[4] Ding Hui, Trajcevski Goce, Scheuermann Peter, Wang, Xiaoyue, Keogh Eamonn, “Querying and mining of time series data:experimental comparison of representations and distance measures”. Proc. VLDB Endow 1 (2): 1542-1552, 2008.

[5] 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.

[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.