

## 時系列データの動向概要を示す自然言語生成手法への一考察

青木花純 †

小林一郎 ‡

† お茶の水女子大学 理学専攻

‡ 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、センシング技術の発達や、IoT への関心の高まりにより、多様な時系列数値データを利用する場面が増えている。時系列データそのものを表示するだけでは、人が内容を理解することは難しく、テキスト表現等を用い、動向概要を示すデータを付与することが多く行われている [1, 2]。それに伴い、時系列数値データから動向概要を示すデータ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、観測された時系列数値データから、自然言語を生成する研究が盛んになっている。本研究では、日経平均株価を例として、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手法を提案し、実験及び考察を行った。

### 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。

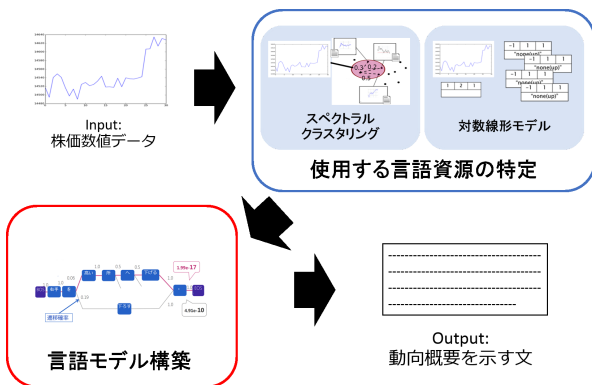


図 1: 概要図

新たに観測された時系列数値データを基に、過去のデータから類似度の高い時系列データを選択する。選択された言語資源には必要に応じて重み付けを行い、バ

イグラムの言語モデルを構築する。このように構築された言語モデルに対して動的計画法を用いることで、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。テキスト生成に使用する言語資源を特定する手法として、対数線形モデルを用いる手法と、時系列データの類似度指標を用いてクラスタリングを行う手法の 2 つを使用した。

#### 2.2 対数線形モデルによる識別器を用いた手法

過去に観測された時系列データには人手でラベル  $r$  を付与し、Symbolic Aggregation approxImation(SAX) 法を用いて 5 分毎に次元圧縮されたデータに変換する。そして新たなデータ  $d$  に対して、そのラベル  $r$  を判定する識別モデルを式 (1) に示すように、対数線形モデルによって構築する。素性ベクトル  $\phi$  は前場、後場における 5 分ごとの圧縮されたデータで構成されるとする。また、 $r$  は動向内容を示すラベルとする。 $Z_{d,w}$  は正規化係数である。

$$P(r|d) = \frac{1}{Z_{d,w}} \exp(\mathbf{w} \cdot \phi(d, r)) \quad (1)$$

過去のデータのうち、判定されたラベルと同じラベルが付与されたデータがもつ言語資源を言語モデル構築に用いる。この手法は人手でラベルを付与するため、コストはかかるが正確な識別器を構築できると考えられる。

#### 2.3 スペクトラルクラスタリングを用いた手法

新たな観測データと過去に観測された時系列データに対してスペクトラルクラスタリング [4] を行い [4]、新たに観測されたデータと同じクラスタに分類された時系列データが持つ言語資源を用いて、言語モデルを構築する。その際、新たに観測された時系列データとの類似度によって各言語資源に重み付けを行う。類似度には Dynamic Time Warping(DTW) 距離を用いた [5]。この手法は人手によるコストがかからないが、類似度計算における計算コストが大きい。しかし、動的に言語モデルを構築するので、識別器による言語モデルよりも柔軟性が高く多様なテキストを生成できると考えられる。

A Study on Natural Language Generation Method Describing the Trends of Time Series Data  
 †Kasumi AOKI ‡Ichiro KOBAYASHI  
 ‡‡Ochanomizu university

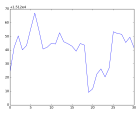
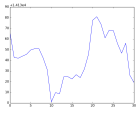
株価動向	識別器による生成文	クラスタリングによる生成文
	正解文：一時上げ幅を拡大した。 ● 上げ幅, を, 拡大, し, た, 。, …, EOS	● 一時, 上げ幅, を, 拡大, し, た, 。, …, EOS
	正解文：方向感の定まらない動きとなった。 ● 下げ, 幅, を, 拡大, し, た, 。, …, EOS	● 下げ, 幅, を, 拡大, し, た, 。, …, 後, 一時, 下げ, 幅, を, 拡大, し, た, 。, …, EOS

表 1: 言語モデルによって生成された文

## 2.4 言語モデルの構築と文生成

選択された言語資源を用いてバイグラムモデルを構築し確率的に尤もらしい単語の組み合わせを見つけることによって、テキストを生成した。バイグラムモデルにおいては、文長が長い文ほど尤度が低くなってしまふ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、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]。

## 3 実験

### 3.1 実験設定

対数線形モデルにおける識別ラベル数およびスペクトラルクラスタリングにおけるクラスタ数は9つとした。ま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、後場の各時間帯にわけて収集した。実験に使用したテキストデータ<sup>\*</sup>、および数値データ<sup>†</sup>は、2013年2月25日~2014年12月30日に収集された451日分の902個のデータを用いた。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 実行結果

提案した手法によって、新たにデータが観測された際、過去に観測されたデータから必要なデータを選択し、言語モデルを構築し、尤もらしい単語の組み合わせを選択する事によってテキストを生成した。使用した識別器の識別精度は約37.2%であった。また、生成された文の例を与えられた時系列データおよびテキストデータと共に表1に示す。

### 3.3 考察

生成されたテキストは短文が多かった。特に識別器を用いる手法を用いた場合は単語数が6以上のテキストは生成されなかった。その理由として、使用した過

去のテキストデータに短文が多くみられたことが挙げられる。また、“下げ幅を拡大した。その後一時下げ幅を拡大した”といった同じ表現を繰り返す冗長な文が生成される場合もあったため、構築する言語モデルはバイグラムだけではなくトライグラムなどを用いるなどさらなる工夫が必要があると考えられる。

## 4 おわりに

本研究では、日経平均株価を例として、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。新たに与えられた時系列数値データに基づいて選択された言語資源を用いて、言語モデルを構築し、動的計画法を用いて尤度の高い単語の組み合わせを得ることで文生成を行い、言語資源の特定には識別器を用いる手法とクラスタリングを用いる方法を用いた。今後の課題として、それぞれの言語資源の特定手法についてクラスタ数などのパラメータも含めたさらなる精査や、生成分の評価手法の検討および生成文の客観的な評価が挙げられる。

## 参考文献

- [1] 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.
- [2] Gkatzia, D., Hastie, H. and Lemon, O., Finding middle ground? Multi-objective Natural Language Generation from time-series data, the 14th European Association for Computational Linguistics, pp.210-214,2014.
- [3] 小林瑞希, 小林一郎, 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, Germany “A Tutorial on Spectral Clustering”, Statics and Computing 17 (4), 2007.
- [5] Ding Hui, Trajcevski Goce, Scheuermann Peter, Wang, Xi-aoyue, Keogh Eamonn, “Querying and mining of time series data:experimental comparison of representations and distance measures”. Proc. VLDB Endow 1 (2): 1542-1552, 2008.

<sup>\*</sup>ADVFN:http://jp.advfn.com/より取得

<sup>†</sup>IBI-Square Stocks:http://www.ibi-square.jp/より取得