言語モデルに基づくテキスト生成を対象とした言語資源の特定手法の検討

青木　花純
お茶の水女子大学大学院 理学専攻 情報科学コース

小林　一郎
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]．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を生成する言語モデルを構築するための言語資源の特定手法について考察する．

2　時系列データからの文生成

2.1　概要

本研究は，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テキスト内容の対応関係を学習することによって，観測された時系列数値データの動向概要を示すテキストを自動生成する言語モデルの構築を目的とし，言語モデルを構築する言語資源の特定のために提案した 2 つの手法を比較し，考察する．図 1 にテキスト生成の概要を示す．

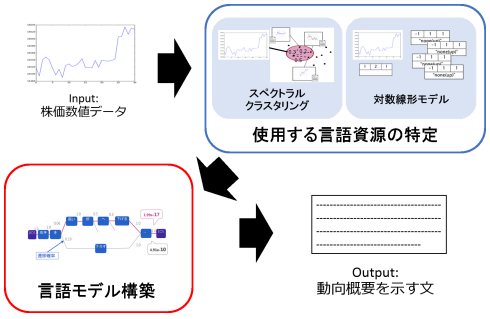


図 1: 研究概要図

1 つ目の手法では，過去に観測された時系列データを学習データとして，新たに時系列数値データが与えられた際に尤もらしいテキスト内容 (以下,「中間表現」と呼ぶ) を決定する識別モデルを構築する．識別モデルの構築には対数線形モデルを用いた，次に識別器によって識別された「中間表現」と同じ「中間表現」がつけられている過去の時系列データと対になっている株価の動向を説明する文書を言語資源として特定し，バイグラムモデルを構築する．

2 つ目の手法では，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 SPC 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を言語資源として特定し，バイグラムモデルを構築する．その際，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．

以上の手法によって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　識別器を用いた言語資源の特定

まず，過去に観測された時系列データに人手で表 1 に示すような「中間表現」を付与し，Symbolic Aggregation approXimation(SAX 法) を用いて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「中間表現」r を判定する識別モデルを式 1 に示す対数線形モデルを用いて構築する．

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

表 1: 中間表現とその意味内容

中間表現	意味内容
UU(UpUp)	上昇し続けた
DD(DownDown)	下落し続けた
SS(StayStay)	あまり変化はなかった
UD(UpDown)	上昇した後，下落した
DU(DownUp)	下落した後，上昇した
DS(DownStay)	下落した後，あまり変化しなくなった
SD(StayDown)	あまり変化していなかったが，その後下落した
US(UpStay)	上昇した後，あまり変化しなくなった
SU(StayUp)	あまり変化していなかったが，その後上昇した

素性ベクトル ϕ 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，$Z_{d,\boldsymbol{w}}$ は正規化係数である．構築された識別器を用いて判定された「中間表現」r_1 と同じ「中間表現」r_1 が付与されている過去の時系列データと対に収集した言語資源として特定する．この手法は人手でデータに「中間表現」をを付与するため，コストはかかるが各中間表現の元で言語資源が管理されている．そのため，安定した言語資源の特定ができるが，生成されるテキストの自由度が少ない傾向にある．

2.3 クラスタリングによる言語資源の特定

2.3.1 時系列データのクラスタリング

時系列データの分類にはスペクトラルクラスタリング (SPC) を用いた．SPC 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 を行っていく事によって，データをクラスタリングする手法である．SPC を用いるにあたって，クラスタ数は予め設定する必要がある．また本研究では，時系列データ同士として類似度に各時系列データの Dynamic Time Warping(DTW) 距離 [6] および Prefix and Suffix-Invariance DTW(ψ -DTW) 距離 [7] を用いた．

- **DTW 距離**

  DTW 距離とは，時系列データの各点の距離を総当りで計算し，距離行列を作成し，その距離行列を元に求めた時系列データ同士の距離が最短となるパスにおける総コストである．時系列データにおける挙動の周期性や挙動の長さが違う場合でも高い類似度を示す距離尺度として用いられている．window 幅などの制約がある．

- **ψ-DTW 距離**

  従来の DTW 距離では実世界の時系列データにおいて適切な類似度を示せない場合がある．それは各時系列の区切りが適切でないことによるものである．その問題を解決するために提案された ψ-DTW は制約パラメータ r を DTW の距離に追加することによって，各時系列データの「Prefix/Suffix」の情報に左右されない類似度指標となっている．

Algorithm 1 ψ-DTW

Input: x,y: sequences, relaxation factor parameter r

Output: ψ-DTW distance.

$M \Leftarrow$ infinity matrix $(n+1, m+1)$

Initialize $M([0,r],j)$ and $M(0,[0,r])$

for all i such that $1 \leq i \leq n$ **do**

  for all j such that $1 \leq j \leq m$ **do**

    $M(i,j) \Leftarrow \mathrm{c}(\mathrm{x}_i, y_i) + min(M(i-1, j-1), M(i, j-1), M(i-1, j))$

  end for

end for

$minX \Leftarrow$min(M([n-r,n],m)),minY$\Leftarrow$min(M(n,[m-r,m]))

return $min(minX, minY)$

2.3.2 特定した言語資源の重み付け

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．

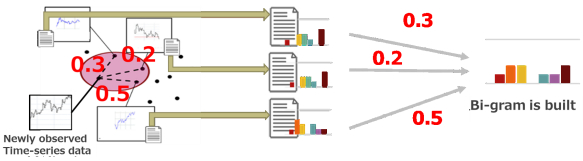
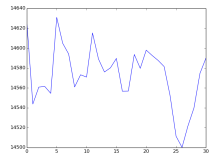


図 2: 重み付けの図

2.4 言語モデルによるテキスト生成

以上の手法を用いて構築した各言語モデルに対して，動的計画法を用いることによって尤度が高くなる単語の組み合わせを獲得し、新たに観測された時系列データの動向概要を示すテキストを生成する．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

表 2: 生成されたテキストの一例

時系列数値データ	正解文	
	一時上げ幅を拡大した	
手法	ラベル/アルゴリズム	生成文
識別器を用いる手法	識別結果：「UU」	上げ幅, を, 拡大, し, た,。, ⋯, EOS
SPC を用いる手法	DTW	上げ幅, を, 拡大, し, た,。, ⋯, EOS
	ψ-DTW	一時, 下げ, 幅, を, 拡大, し, た,。, ⋯, EOS

構築する際には，図 3 のよう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．

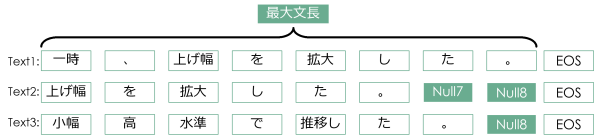


図 3: 仮想単語 null の挿入

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

対数線形モデルにおける識別ラベル数および SPC におけるクラスタ数は 3 もしくは 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書は前場，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 [1]，および数値データ [2] 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．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テキストを生成した．

[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 実行結果および考察

実行結果の例として，提案した 2 つの手法を用いて生成されたテキストを時系列数値データおよび正解テキストとともに表 2 に示す．2 つの手法において生成されたテキストには短文が多くみられた．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．また意味的に冗長なテキストが生成されてしまうことがあるため，過去に観測されたテキストのバイグラムのみを学習させるだけではなく，形容詞や動詞などの重要単語の組み合わせと文法を組み合わせる手法を用いた言語モデルの構築なども必要であると考えられる．生成された文の例のように，識別器を用いる手法によって生成されたテキストよりもクラスタリングを用いる手法によって生成されたテキストの方が自由度が高い文章が生成されている．一方で，SPC を用いる方法ではクラスタリングを予め設定する必要があり，設定したクラスタ数は言語資源の特定に大きな影響を与えると考えられる．そのため，今後はクラスタ数も推定するノンパラメトリックなクラスタリング手法が必要であると考える．

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テキスト生成を行い，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，クラスタリングを用いる手法においてクラスタ数を自動推定する手法の選択などが挙げられる．

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

[7] Silva, D. F., Batista G. E. A. P. A. and Eamonn Keogh, " Prefix and Suffix Invariant Dynamic Time Warping ", in Proceedings of the IEEE International Conference on Data Mining, 2016.