言語モデルを用いた株価の動向を記述するテキスト生成への取組み

青木 花純　　小林 一郎
お茶の水女子大学 理学部 情報科学科
{g1120501,koba}@is.ocha.ac.jp

1 はじめに

近年，センサや処理能力の高い計算機の発展により膨大なデータを処理することが可能になってきた．センサなどから観測されるデータは，時系列数値データであり，様々な用途で利用される場面が増えている．一方で，そのような数値データの概要を人が把握するには，そのままの表示では困難となる．そのため，テキスト表現等を用いた概要を同時に示す事も多く，株価や為替の動向などを示す新聞記事などには，グラフ化された時系列データと共にその動向を説明するテキストも記されている．また，視覚情報として観測されるデータも時系列数値データとして処理され，自然言語処理研究の分野においても，近年，視覚情報や数値情報など非言語情報を言語情報として表現するテキスト生成の研究が盛んになっている [1, 2, 4]．本研究では，時系列数値データの代表的かつ身近な例である，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む．

2 時系列データからのテキスト生成

2.1 概要

本研究では，過去の株価動向のパターンとそれを説明する文章内容の対応関係を学習し，文章内容を表現するのに必要な言語資源を利用することによりテキストを生成する．まず初めに，観測された株価の数値データを教師データとして，新たな数値データが与えられた際に，尤もらしい文章内容（以下「中間表現」と呼ぶ）を決定する識別モデルを構築する．識別モデル構築には対数線形モデルを用いる．次に，収集した株価の動向を説明する文章から，中間表現ごとにバイグラムモデルを構築し，識別された中間表現の下，使用されるバイグラムモデルを動的計画法で解く事により，その内容を示す尤もらしい単語の組合わせを抽出することでテキストを生成する．

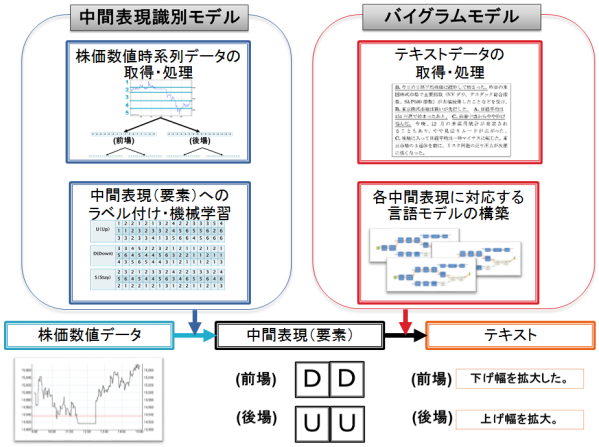

図 1: 時系列データからのテキスト生成の概要

2.2 株価データ

- **数値データ：**
  株価の数値データとして，日経平均の 5 分足データを用いる．前場と後場の 2 つの区間に分け，それぞれ 2 時間 30 分の離散時系列数値データを扱う．データは株価や先物，株価指数のデータを配布しているインターネットサイトである株価データダウンロードサイト [1] から収集した．

- **テキストデータ：**
  日経平均株価の数値データと共に配信されている，その動向を説明するテキストを用いる．データは株価や先物，またそれらに関連するニュースを取り上げている ADVFN[2] から収集した．テキ

[1] http://k-db.com/
[2] http://jp.advfn.com/

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

ストデータは，事前に数値データとの共起関係により対応づけを行う．

2.3 中間表現

中間表現は，時系列数値データの動向およびそれを説明するテキストの「意味内容」に相当し，両者の橋渡しを行うものである．中間表現は，前場，後場の各時間帯を 2 つに分けて捉えている．以下に，本研究で使用した中間表現を示す．

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

中間表現によって表される意味内容は，数値による動向とテキストによる説明が，前場と後場のそれぞれにおよそ 2〜3 つ観測される事実に基づき，表 1 のように設定した．

3 時系列データ処理

3.1 SAX 法

時系列数値データからその動向をパターンとして抽出するために，本研究では，SAX（Symbolic Aggregation approXimation）法 [3] を用いた次元圧縮法を用いて、データを圧縮させる．まず，正規分布に従って株価を 6 つの範囲に分け，各範囲に文字を割り当て，株価の 5 分毎の高値を対象に観測した数値を文字に変換する．5 分足データを利用しているため，変換された文字は，前場，後場ともに各 31 個となる．

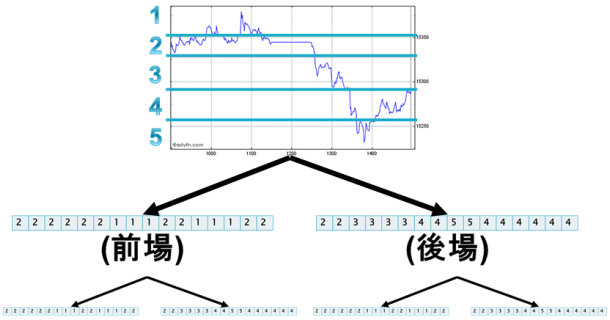

図 2: SAX 法による次元圧縮

3.2 時系列データ動向識別モデルの生成

SAX により次元圧縮されたデータ d に対して，その内容 r を判定する識別モデルを対数線形モデルを用いて構築する．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，r は中間表現を構成する要素（例,「U」等）とする．$Z_{d,\boldsymbol{w}}$ は正規化係数である．

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

4 言語モデルによるテキスト生成

4.1 収集テキストからの言語モデル構築

収集した株価の概況テキストの例を以下に示す．

表 2: 株価の概況例

B. 今日の日経平均株価は続伸して始まった。昨日の米国株式市場で主要指数（NY ダウ、ナスダック総合指数、S&P500 指数）が大幅続伸したことなどを受け、**B.** 東京株式市場は買いが先行した。 **A.** 日経平均は 151 円高で始まったあと、 **C.** 前場中頃からやや伸び悩んだ。 今晩、12 月の米雇用統計が発表されることもあり、やや見送りムードが広がった。**C.** 後場に入って日経平均は一時マイナスに転じた。東京市場の 3 連休を前に、リスク回避の売り圧力が次第に強くなった。

テキスト生成の対象となる表現は下線で示された部分である．その内容を表 3 に示す．

表 3 に示すように，収集テキストから生成対象となる文を抽出し，それに基づき言語モデルを構築する．テキスト生成の対象となる情報は，当日の時系列データの動向や前日のデータとの比較に言及し，株価の変動理由や，関連する時事ニュース等は本研究では対象

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

表 3: テキスト生成内容

識別記号	テキスト生成内容
A	取引開始直後の株価と前日終値との比較
B	取引開始直後の動向
C	取引時間中の動向（前場/後場）

外とする．また，表 1 における中間表現は，**A**，**B**，**C** の内，**C** にのみ対応している．

4.2 バイグラムモデルによるテキスト生成

本研究では，バイグラムモデルを用いたテキスト生成を行う．株価動向を示す中間表現毎にテキスト生成のためのバイグラムモデルを構築するため，各中間表現の意味内容を表すテキストを収集し，言語モデルを構築する．これにより，観測された時系列データを識別する中間表現が選択されると，それに対応するバイグラムモデルが選択され，テキスト生成が行われる．

しかし，同じ現象に対しても人によって説明の仕方は様々であり，選択する表現や文の長さが異なる．一般に，構築したバイグラムモデルから尤度の高い単語の組合わせを抽出することによってテキスト生成を行う場合，単語数が少ない文のほうが尤度が高くなってしまう．このことから，本研究では，文長に左右されないテキスト生成を行うために，小林ら [4] によって提案されている，疑似単語として番号付き null ラベルをバイグラムモデルの中に入れることにより，テキスト生成を行う．番号付き null ラベルは，中間表現に対する言語モデルを構築する際に予め収集したテキストデータの内，最大文長の文の単語数を調べ，その語数に満たない収集文の末尾に追加される．それらは，単語として同等に扱われ，言語モデルを構成する資源の一部となり，生成された短い文に対しても，尤度の次数を一定に保った下での生成文の順位づけが可能となる．番号付き null ラベルを取り入れたバイグラムモデルを動的計画法を用いて探索することにより，尤もらしい単語の組合わせからなる文の生成を行う．

5 実験

本章では，上記に説明した手法を用いて，新たな株価時系列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．

5.1 実験設定

テキストの生成内容は，表 3 に示された **C** に限定し，表 1 に示された各中間表現に対してテキストを生成する．言語モデルを構築するためのテキストは中間表現に対応づけて収集され，前場，後場の各時間帯において 2 つの動向を含むものを対象とする．実験に使用したテキストデータは，2014 年 8 月 12 日〜2015 年 1 月 9 日に収集された 100 日分のデータである．また数値データは 2014 年 10 月 14 日〜2015 年 1 月 9 日に収集された 53 日分のデータである．収集したテキストデータの特徴を表 4 に示す．

表 4: テキストデータの特徴

中間表現	文数	語数	最大文長
UU(UpUp)	31	225	20
SS(StayStay)	39	353	16
DD(DownDown)	23	155	11
US(UpStay)	31	356	25
SU(StayUp)	14	187	24
SD(StayDown)	11	114	21
DS(DownStay)	22	275	28
DU(DownUp)	5	61	18
UD(UpDown)	5	83	25

5.2 実験結果

- **中間表現識別結果**
  中間表現判別のために，中間表現の要素を (1) を適用した対数線形モデルを適用し，判別した．この識別器は 212 のデータを 10 個のデータセットに分割し，10 分割交差検定を行った結果，中間表現の要素においては約 47%，中間表現においては約 26%の精度を示した．

- **バイグラムモデルによるテキスト生成結果**
  各中間表現に対して最大文長を考慮したバイグラムモデルを構築し，動的計画法を適用することで，株価データの概要を説明する尤もらしい文を生成した．例として，2 つの中間表現に対して最も尤度の高かった上位 3 文を表 5 に示す．

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.

表 5: 生成された文（上位 3 件）

中間表現	生成文	尤度
US	やや，伸び悩ん，だ，水準，で，推移，し，た，。，null10，null11，null12，null13，null14，null15，null16，null17，null18，null19，null20，null21	7.19e-43
	やや，伸び悩ん，だ，水準，で，推移，し，た，。null9，null10，null11，null12，null13，null14，null15，null16，null17，null18，null19，null20	3.09e-43
	上げ，幅，を，拡大，し，た，が，、，次第に，伸び悩ん，だ，。，null10，null11，null12，null13，null14，null15，null16，null17，null18，null19，null20，null21，null22，null23	9.66e-45
DU	プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，転じ，た，。，EOS	2.83e-28
	プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，	2.72e-28
	下げ渋り，、，プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，転じ，、，プラス，に，転じた，	2.38e-28

5.3 考察

実験結果から，中間表現を識別する対数線形モデルの精度が低いことが確認された．中間表現の要素の個数を 2 にしたこと，要素を判別する際のデータの区切りが不適切であることが理由であると考えられる．
また，各中間表現に対するバイグラムモデルにおいて，各中間表現の「意味内容」に適切な文を生成できた中間表現（例,「US」）が見られる一方で，特定の単語組を繰り返してしまい，適切ではない文を生成した中間表現（例,「DU」）も見られた．バイグラムモデルに統語規則を組み込むなどして改善出来るのではないかと考えている．

6 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いくつかの次元圧縮手法を適用することによって，そのパターンを抽出し，対数線形モデルにより中間表現との対応関係を識別するモデルを構築した．また，時系列数値データを説明するテキストデータを収集し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尤度の高い単語の組み合わせを得ることにより文生成を行った．その際，文長に対する処理として，言語モデルを構築する際に収集文の最大文長（最大単語数）に生成する文の長さを合わせ，満たない文に番号付き null ラベルを導入し，それらを単語と同様として見なすことで，尤度に基づく文生成において，単語数の制限を受けない自然言語文の生成を可能にした．

上記，処理において構築した中間表現識別モデルの識別精度は低い結果となった．これは，現時点での時系列データの次元圧縮手法において適切な処理が出来ていないためと考えられる．これに関して，前場・後場の各時間帯に中間表現の要素数の制限を設けないように数値データを圧縮する手法について考察を深めるつもりである．また，テキスト生成において，本研究では統語的知識を取り入れてはいない．今後は，統語的知識を取り入れると共に，現在，個々の中間表現に対して対応するバイグラムモデルを準備している点を改善していきたいと考える．

参考文献

[1] Haonan Yu and Jeffrey Mark Siskind, Grounded Language Learning from Video Described with Sentences, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 53-63, Sofia, Bulgaria, August 4-9 2013.

[2] M.Regneri, M.Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal, Grounding Action Descriptions in Videos, Transactions of the Association for Computational Linguistics (TACL), 2013.

[3] Lin, J., Keogh, E., Lonardi, S. and Chiu, B., A Symbolic Representation of Time Series, with Implications for Streaming Algorithms DMKD ’03, 2003.

[4] 小林瑞季, 小林一郎, 麻生英樹, 動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組み, 第 27 回人工知能学会全国大会, 2D5-OS-03b-3, 2013.

Copyright(C) 2015 The Association for Natural Language Processing.
All Rights Reserved.