3B3-NFC-4-2

時系列データの言語化への取り組み

An Approach to Verbalizing Time-series Data

小林一郎[*1]
Ichiro Kobayashi

[*1]お茶の水女子大学大学院人間文化創成科学研究科理学専攻
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu Univeristy

Most data observed in our lives are time series data. So, we need a method to be able to access and easily utilize the data. As a representative method to realize it, visualization is widely used. On the other hand, as we see that there are many texts, e.g., newspaper articles reporting the trends on stock prices, foreign exchange rates, weather conditions, etc. Explaining time-series data with words is also widely used. With this background, in this study, we focus on explaining time series data with words and propose a method to verbalize time series data from a macroscopic viewpoint – which is that we aim to verbalize time series data by visually recognizing the shapes of a line chart of time series data. We apply our proposed method to verbalize stock price time series data and evaluate the results of generated texts.

1. はじめに

観測されるデータの多くは，時系列データであり，アンビエントコンピューティング，ユビキタスコンピューティングなど環境知能に関する研究が発展すると共に，ライフログ取得に関する研究も進められ，現在では多くの時系列データを観測する手法や環境が開発されてきている．

一方，観測した大量の時系列データに内在する情報や知識を解釈する必要性も生じている．このような大量の時系列データを人が効果的に解釈するためには，データの可視化などの方法があるが，可視化が常に適用可能と言えない場合も存在する．例えば，可視化ができない機器を使って時系列データの動向を伝えなくてはならないときや，目が不自由な人へその動向を伝えたいときなどがある．このような場合は，時系列データを自然言語で表現し，その動向を伝えられるということはとても大きな利点となる．また，医療行為の場面において観測された生体情報データを可視化したものよりも言語で説明した文章として出力されたものの方が適切な処置を施せるとの報告 [1, 2] や携帯電話のマニュアルにおいてテキストと図示による両方の要約を比較した心理学研究においてテキストによって情報を提示された方が意思決定が容易であるという結果なども得られている [3]．このようなニーズから時系列データの言語化において様々な手法が提案されているが，本研究では，時系列データの概観を捉える巨視的な観点から言語化を行う手法の提案を行い，対象時系列データとして日経平均株価を取り上げ，提案手法により言語化されたテキストと実際の新聞記事とを比較することによって提案手法の考察を行う．

2. 関連研究

数値データを言語で表現する研究は古くからされている．1983 年には，Kukich [4] によって，Wall Street Journal の中の記事のような株価市場に関する日々の報告をするシステムを開発した．1994 年には，Goldberg ら [5] によって，気象データを元に 2ヶ国語の気象報告を生成するシステムが開発されている．Kukich や Goldberg らのシステムでは，どのような内容を文章として出力するかよりもむしろ談話構造や出力文章の文法に焦点が当てられていた．一般的に，数値データを時系列データとしての側面からその内容を文章化する際には，文章化する内容の決め方がより難しくなる．Boyd[6] は，時系列データの解析に波形解析に利用されるウェーブレット解析を導入しており，実際に気象報告を対象にしたテキストを生成して提案手法の検証を行い，他の時系列データにも手法の適用が可能であると報告している．

長期に渡り進められている時系列データ言語化の研究プロジェクトとしては，英国 Aberdeen 大学の研究チームによる種々の時系列データモデル化の研究プロジェクトである SumTime プロジェクトが挙げられる [7, 8, 9]．SumTime プロジェクトは，時系列データの分析と自然言語生成技術を融合させた時系列データの言語（英語）による要約生成のより良い技術を開発することを目的として進められ，様々なドメインを対象にシステムが構築された．SumTime-Mousam[7, 10] では，気象データを対象にしたテキスト生成をおこなっており，気象予報官がより質の高い気象報告を作成することを支援する．また，その枠組みをガスタービンの大量のセンサデータを文書として要約するのに適用した SumTime-Turbine プロジェクトも行われた [11, 12]．SumTime-NEONATE[8, 13] は，新生児集中治療室 (NICU:Neonatal Intensive Care Unit) で観測される多量の生体信号をテキストとして要約し報告することによって，医療スタッフの意思決定支援を支援すること，および，新生児の健康状態をテキストとして患者の家族にその状態をわかりやすく伝えることを目的として開始され，その後，Baby Talk プロジェクト [9, 14, 15] として，現在，同チームの主要な研究プロジェクトの一つとして進められている．

著者は，日経平均株価を対象として，時系列データを人が視覚的に判断するのと同じようにグラフの形状を捉えてその状態を言語で表現する手法を提案する．

3. システム構成

最もよく知られている一般的な自然言語生成システムのアーキテクチュアは，Reiter& Dale[18] によって示されている 3 つのステージが橋渡しになっているモデル (Pipe-line model) で

連絡先: 小林一郎，お茶の水女子大学大学院人間文化創成科学研究科理学専攻，〒 112-8610 東京都文京区大塚 2-1-1，koba@is.ocha.ac.jp

あり,「文書プランニング」,「マイクロプランニング」,「実体化」の 3 つのステージにより構成されている．自然言語生成システムのアーキテクチュアは，どのようにモージュール化するかによって多様に表現されることが指摘されている [19] が，時系列データを入力情報とするほとんどのアーキテクチュアには，新たに「データ分析」,「データ解釈」といった 2 つのステージが，文書プランニングステージの前へ追加される*1（図 1 参照）[20, 21].

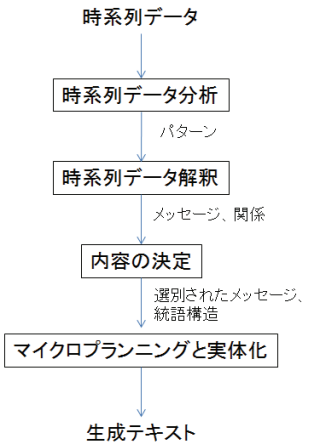


図 1: 時系列データを入力とする自然言語生成システムの一般的なアーキテクチュア

このアーキテクチュアに従い，本研究における言語化システムの各ステージでの処理を説明していく．

3.1 時系列データの分析手法

本研究では，日経平均株価の 10 分足データの動向をグラフの形状として捉えるために，線形最小二乗法により時系列データのグラフを近似する曲線を求め，その近似多項式曲線の振る舞いを観測することにより時系列データを分析した．近似曲線は，株式市場の前場と後場それぞれに適用し 4 次多項式で表現した．多項式の次数は小さすぎるとグラフの挙動を表現しきれず，大き過ぎると余分な挙動まで表現してしまう．グラフの振る舞いを言葉で表現する際に，この近似曲線の部分形状とそれを表現する語彙が適切に一致する必要があるため，収集した実際のコーパス（日経平均株価動向を解説する新聞記事）においてグラフの挙動を説明する語彙がどのようなグラフの形状を指しているかを分析し，その結果に基づいて多項式の次数を決定した．

解説記事で使用される言語表現は，この 4 次多項式が極値の取り方によって表現可能となる 12 のタイプに分類されるグラフの部分形状を表現することになる．このことから，その 12 のタイプに分類されるグラフの部分形状を対象に言語化を行う．図 2 にタイプごとに分類された部分形状の一部を示す.

3.2 データの解釈

日経平均株価の一日の動向を解説した実際のコーパスから時系列データの動向を表す線形最小二乗法によって得られた近似グラフの部分形状や値の変化を示している言語表現を抽出し，図 3 に示すグラフの形状の数式的な解釈とそれを表現する言語を対応づけた辞書の構築を行った.

*1 同じ SumTime プロジェクトの枠組みでも SumTime-Turbine [11, 12] においては，システムを 9 つのモジュールに分割したアーキテクチュアを採用している.

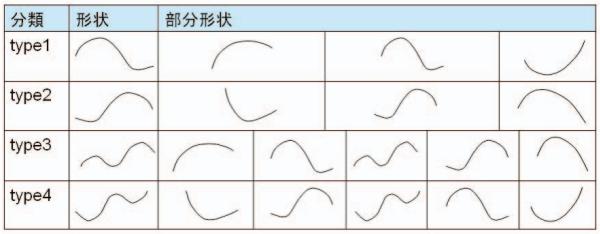


図 2: タイプごとに分類された部分形状（一部）

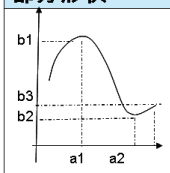
部分形状	短文+時間帯	特徴
	売りが優勢だった	|b2-b1|/|MAX-MIN|>0.4 |a1-a2|/|max-min|<0.7
	売りが広がった	|a1-a2|/|max-min|>0.7
	売りが優勢になる場面があった	|b2-b1|/|MAX-MIN|>0.4 |b2-b3|/|b2-b1|>0.5 |a1-a2|/|max-min|<0.7

図 3: グラフの形状認識と言語表現

本研究における時系列データの解釈の方針は，統計的な特徴量に基づいてグラフの特徴を捉えるのではなく，人がグラフ化された時系列データを巨視的に捉え，自らが持つ言葉によってグラフ化された時系列データの部分的な形状を言い表すことにより，解釈が為されるというものである．そのため，認識されるグラフの部分形状は語彙の立場から定義される．グラフの部分形状を認識するための語彙は，実際の株価動向を説明する新聞記事より収集して構築されている.

3.3 内容の決定

内容の決定は，どの情報をテキストの中に包含し，どの情報を構造化（節にする）すべきかを決定する．本研究では，数値による値を直接言及するタイプのテキスト（「タイプ 1 テキスト」と呼ぶ）とグラフの部分形状を言及するタイプのテキスト（「タイプ 2 テキスト」と呼ぶ）の 2 種類のテキストを生成している．タイプ 1 テキストにおいては，その内容は株価の動向を伝えるのに必要な情報となる，過去の始値，終値，高値，安値の数値情報となり，タイプ 2 テキストにおいては，システムが認識したグラフの部分形状となる．認識されたグラフの部分形状は観測された時間順序に従い，テキストとして生成される.

3.4 マイクロプランニングと実体化

マイクロプランニングと実体化（realization）は，文書内容として決定されたものを最終的に出力結果となる文書に変換する.

タイプ 1 テキストの生成には，3 種類のテンプレートを用いて，空欄に観測した値を入れることによりテキストを生成する．テンプレートのひとつの例を以下に示す.（下線部の空欄に適切な値や表現が入る.）

> 「__日の東京株式市場で日経平均株価は__．終値は__円__銭__（__%）__の__万__円__銭で、__円台を回復した.」

タイプ 2 テキストは，株式市場が前場と後場の二つの時間

帯において取引がされていることを考慮し，前場，後場を近似多項式にてグラフの形状を認識する．グラフの部分形状を定量的に解釈し語彙に変換する辞書（図 3 参照）を経て，各部分形状に対して語彙表現を付与する．状況語（特に「時間」を言及するもの）を付与し，最後に，因果関係などを示す接続詞を付与することによりテキストが生成される．グラフの部分形状から生成されるテキストの例を以下に示す.（下線を引いた語彙表現はグラフの部分形状を認識したものに相当する）

「28 日前場中ごろの東京株式市場で日経平均株価は 上昇に転じる 場面があった．相場は 堅調に推移した が，大引け間際に まとまった売り が出て 上げ幅を縮小した.」

3.4.1 語彙表現

出力として生成されるテキスト内に使用される語彙を適切に選択することも重要な課題となる．本研究においては，株価動向を説明する実際のコーパスとそれに対応する株価のグラフの部分形状の対応関係を観測することにより辞書を構築している．辞書構築にあたっては，コーパスを分析することにより，グラフの部分形状を適切に表現する語彙，文を収集することにより構築を行っている．

現時点において，辞書内には，部分形状を表現できる短文が 64 種類 (例：「売りが広がった」,「じり高歩調となった」,「反発」)，時間帯が 9 種類 (例：「前場」,「大引けで」)，接続詞が 4 種類 (例：「そして」,「なので」) 登録されている．

3.4.2 文法

タイプ 2 テキストは，短文，時間帯，接続詞の実際のコーパスを真似た適切な語彙組み合わせ規則により生成される．その例を以下に示す.

- 時間帯によって先頭に「前場は」,「後場は」をつける.
- 部分形状によっては，時間帯によって「中ごろ過ぎにかけて」,「中ごろに」，などが短文の前につけられる.
- 上げ幅，下げ幅について言及している短文はその前に接続詞「そして」が前につけられる.

4. 言語化とその評価

2009 年 10 月 14 日の株価データより，システムによって実際に生成されたテキストの一例を以下に示す.

- タイプ 1 テキスト：
  「14 日の東京株式市場で日経平均株価は 4 営業日ぶりに反落．物価は前日比 10 円 00 銭（0.1 %）安の 10070 円 00 銭だった．10100 円近くまで上昇する場面があった.」
- タイプ 2 テキスト：
  「前場は，下げに転じる場面もあった．中ごろ過ぎから切り返した.」

上述したように，生成される 2 つのタイプのテキストは，時系列データの特徴を反映したものになっている.

しかし，生成されたテキストにおいては，その妥当性を検証することが難しい．実際のコーパスを正解とし，生成されたテキストとの一致により評価することも可能ではあるが，実際のコーパスも人の主観に基づき生成されたものであることから時系列データを唯一に言語化したものとは言えない.

先行研究においても同様の問題に対処するために様々な取組がなされてきた．SumTime の枠組みによって生成されたテキストに対する初期の段階の評価として，生成されたテキストの長さやデータの言語化される箇所の特定方法などに対して評価がなされた [12, 22]．その後，情報をわかりやすく伝える基準としてグライスの公準（Gricean maxims）[23] を取り入れ，生成されたテキストにおいて，伝えるべき情報の伝達と，不要な情報伝達の回避という観点からの評価を取り入れている [24]．また，SumTime によって生成されたテキストを最終的にユーザに提示するために，人が後編集にかける労力の観点からも評価を行っている [25].

本研究においては，生成されたテキストについて，同日の株価動向に関する実際のニュースにおいてグラフの形状（株価の動向）を表現するのに使用された表現と (i) 完全に一致するか否か，また，(ii) グラフの形状を捉える数式定義（図 3 の「特徴」項目を参照）において，観測されたグラフの形状を表現する数式のパラメータの包含関係に基づく言語表現の同意（正確には「含意」となる）の検証を行った．このパラメータの値による包含関係に基づく言語表現の同意の例を図 4 に示す.

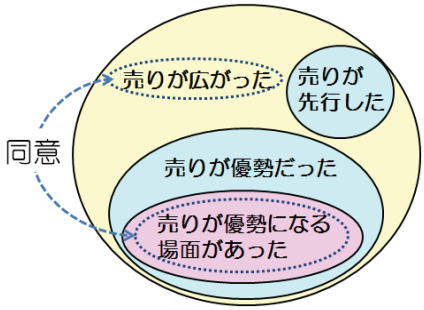


図 4: パラメータの包含関係に基づく言語表現の同意

図 4 においては，言語表現「売りが優勢になる場面があった」を表すグラフの形状を捉える数式のパラメータの値は，言語表現「売りが広がった」のパラメータの値に包含される関係にあるため，生成されたテキストと実際のテキストは同意であるとして，一致するものと評価する．さらに，(iii) ある時系列データに対して，実際のコーパスと較べた際に，その挙動を説明するのに十分なテキストが生成できているかを，グラフの形状を表す表現が適切である生成されたテキストの数から判定する．これら 3 つの評価基準を導入することにより，開発したシステムが時系列データの挙動を適切に説明するテキストを生成できているかについて評価を行った [17]．表 1 は，ある一日の株価動向をテキストとして生成した結果を示している．これによると，生成されたテキストの実際のコーパスに対する一致性は，同意の表現も含めて，およそ 86%となり，これは実際のコーパス生成者（人）の見方に照らした生成能力の観点からのシステムのテキスト生成能力の評価に相当し，グラフの形状認識のために作成した辞書が適切であること，および，人と認識する部分グラフの箇所が同じであることを示す．一方，グラフの挙動に対する一致は，生成されたテキストが正しくグラフの挙動を説明していると見なせるものの数が示されており，実際のコーパスとの一致はなかったとしても，このことからグラフの挙動を説明するのに十分なテキスト数が生成されていると評価できる.

5. おわりに

時系列データの言語化システムにおいては，時系列データの処理・解釈を行い，その結果を用いてテキスト生成を行うと

表 1: 株価動向言語化の結果

グラフ特徴	実際のコーパス	コーパスに対する一致		グラフの挙動に対する一致
		完全	同意	
状態	4	1	2	11
変化率	25	5	23	12
変動量	6	1	3	11
その他	16	3	16	13
合計	51	10	44	47

いう処理の流れは，ほぼ全てのシステムに同様に見られる．それぞれのシステムにおいては，時系列データの性質または言語化の目的に応じて，時系列データに対して適切な分析手法および解釈手段が用意されている．テキスト生成においては，談話構造，統語構造などを対象にしたいわゆる深い処理を行うものは多くはない．これは，この研究の性質として，時系列データを適切に処理するという大きな課題にも研究の重点を置いているということ，および，それほど複雑なテキストの生成が必要とされないこともその理由と考えられる．一方，現状において，状況を考慮した適切な語彙や構文の選択がなされているとは言えないことも今後の課題として挙げられる．また，生成されたテキストの評価については，どのようにも言語で表現することが可能な時系列データを，ユーザの希望に応じて適切にその内容を伝えることができるかという点が重要になる．そのため，時系列データを取得した対象のドメインごとに適切な評価方法が考慮されるべきであると考える．

参考文献

[1] Law, A.S, Y. Freer, J. R.W. Hunter, R.H. Logie, N. McIntosh, J. Quinn, A Comparison of Graphical and Textual Presentations of Time Series Data to Support Medical Decision Making in the Neonatal Intensive Care Unit. Journal of Clinical Monitoring and Computing 19, pp. 183–194, 2005.

[2] van der Meulen, Marian, Robert H. Logie, Yvonne Freer, Cindy Sykes, Neil McIntosh and Jim Hunter, When a Graph is Poorer than 100 Words: A Comparison of Computerised Natural Language Generation, Human Generated Descriptions and Graphical Displays in Neonatal Intensive Care. Applied Cognitive Psychology, 2008.

[3] Langan-Fox J,C Platania-Phung and J Waycott, Effects of Advance Organizers, Mental Models and Abilities on Task and Recall Performance Using a Mobile Phone Network, Applied cognitive psychology, 20, pp.1143–1165, 2006.

[4] Kukich, Karen, Knowledge-Based Report Generation: a technique for automatically generating natural language reports from databases, SIGIR '83: Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval, pp.246–250, Bethesda, Maryland, 1983.

[5] Goldberg, Eli and Driedger, Norbert and Kittredge, Richard I., Using Natural-Language Processing to Produce Weather Forecasts, IEEE Expert: Intelligent Systems and Their Applications, Vol. 9, No.2, pp.45–53, 1994.

[6] Boyd, S., TREND: A System for Generating Intelligent descriptions of time-series data, IEEE International Conference on Intelligent Processing Systems, 1998.

[7] http://www.csd.abdn.ac.uk/research/sumtime/

[8] http://www.csd.abdn.ac.uk/research/neonate/

[9] http://www.csd.abdn.ac.uk/research/babytalk/

[10] Sripada Somayajulu, Ehud Reiter and Ian Davy, SumTime-Mousam: Configurable Marine Weather Forecast Generator. Expert Update6(3):4-10, 2003.

[11] Jin Yu, Jim Hunter, Ehud Reiter and Somayajulu Sripada, Recognising Visual Patterns to Communicate Gas Turbine Time-Series Data. In: Macintosh, A., Ellis, R. and Coenen, F. (ed) Proceedings of ES2002, pp. 105-118, 2002.

[12] Jin Yu and Ehud Reiter and Jim Hunter and Somayajulu Sripada, SumTime-Turbine: A Knowledge-Based System to Communicate Gas Turbine Time-Series Data, The 16th International Conference on Industrial & Engineering Applications of Artificial Intelligence and Expert Systems,pp.23–26, 2003.

[13] Somayajulu G. Sripada, Ehud Reiter, Jim Hunter and Jin Yu, Summarizing Neonatal Time Series Data. In Proceedings of the research note sessions of the EACL03, pp. 167-170, Budapest, Hungary, 2003.

[14] van der Meulen, Marian, Cynthia Sykes, Yvonne Freer, Jim Hunter, Neil McIntosh and Robert Logie, Generating Textual Summaries of Clinical Data: The Baby Talk Project. Archives of Disease in Childhood: Fetal and Neonatal Edition 93, 2008.

[15] Gatt, Albert, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur and Somayajulu Sripada, From data to text in the Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. AI Communications 22, pp. 153–186, 2009.

[16] 小林一郎, 渡邉千明, 奥村奈穂子：グラフとテキストの協調による知的情報処理手法－日経平均株価テキストとグラフの掲示を例にして－, 情報処理学会論文誌,「インタラクション技術の原理と応用」特集号, Vol.48, No.3, pp.1058-1070, 2007.

[17] 関亜沙美，小林一郎：時系列データ言語化への取り組み－日経平均株価を例として－, 第 24 回人工知能学会全国大会,2J2-NFC2-1,2010.

[18] Ehud Reiter and Robert Dale, Building Natural Language Generation Systems, Cambridge University Press, 2000.

[19] Chris Mellish, Donia Scott, Lynn Cahill, Daniel Paiva, Roger Evans, and Mike Reape, A reference architecture for natural language generation systems. Natural Language Engineering, 12:1-34, 2006.

[20] Yu J., E. Reiter, J.R.W. Hunter, and S. Sripada, A New Architecture for Summarising Time Series Data. Proceedings of INLG-04 Poster Session , pp. 47–50, 2004.

[21] Reiter, Ehud, An architecture for Data-To-Text systems. In: Busemann, Stephan (Ed.), Proceedings of the 11th European Workshop on Natural Language Generation , pp. 97–104, 2007.

[22] Somayajulu G. Sripada, Ehud Reiter, Jim Hunter and Jin Yu, Segmenting Time Series for Weather Forecasting. In: Macintosh, A., Ellis, R. and Coenen, F. (ed) Proceedings of ES2002, pp. 193-206, 2002.

[23] Grice, H.P. "Logic and Conversation", In Cole P. and Morgan J. (Eds.) Syntax and Semantics: Vol.3, Speech Acts. Academic Press, New York, pp.43-58, 1975.

[24] Somayajulu G. Sripada and Ehud Reiter and Jim Hunter and Jin Yu, "Generating English Summaries of Time Series Data Using the Gricean Maxims", In Proc. KDD '03, pp.187–196,ACM Press, 2003.

[25] S. Sripada, E. Reiter, and L. Hawizy, Evaluating an NLG system using post-editing. Technical Report AUCS/TR0402, Department of Computing Science, University of Aberdeen, 2004.