日経平均株価を対象にした時系列データの言語化

関亜沙美　(指導教員：小林一郎)

1　はじめに

我々の周囲で観測されるデータの多くは時系列データである．その時系列データの解釈へのアプローチとしてグラフなどのモダリティに表現を変更する可視化などの手法がある．一方，株価や為替の一日の動向などを示すテキストが新聞や WEB ページに掲載されているように，時系列データの振る舞いを言葉で説明するニーズも数多く存在する．そこで，本研究では，時系列データの振る舞いを言葉で説明することに着目し，日経平均株価の動向を例とした時系列データの言語化手法を提案し，システムを開発することを目的とする．

2　日経平均株価テキスト生成システム

2.1　システム概要

先行研究 [1] において開発された「日経平均株価テキスト生成システム」の概要を図 1 に示す．

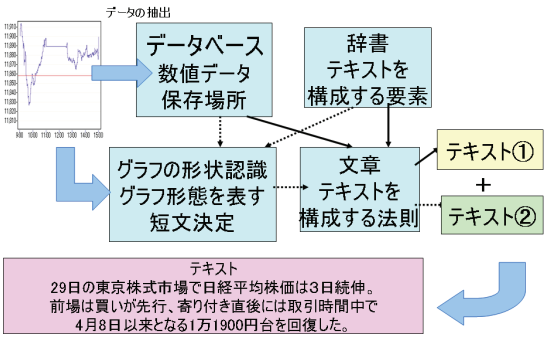

図 1: システムの概要

このシステムによって生成されるテキストは以下の 2 つのテキストタイプに分類され，タイプごとにテキスト生成の処理の流れが異なる．

テキスト①:グラフの形状を踏まえることなしに，データベースからの情報のみから生成できるテキスト．

テキスト②：グラフの形状を踏まえて，かつデータベースからの情報から生成できるテキスト．

本研究においては，テキスト②の自動生成に着目し，テキスト生成の性能向上およびその評価を行う．以下にシステム各部の説明，および，テキスト②の生成処理の流れを示す．

2.2　グラフの形状認識

グラフの動向を把握するとき，グラフが「下がって、上がっている」などの形状によって認識される．グラフを視覚的に把握するために，本研究では，線形最小二乗法を用いてグラフの近似曲線を作り，その近似曲線の振る舞いを捉えることにより，グラフの動向を言語で表す．近似曲線は 5 次多項式で表現されており，この多項式の次数は，グラフの形状を表現している語彙の実際のコーパス (約 1ヶ月分の日経平均株価動向の解説記事) を分析することにより，その最適な次数を 5 次と導いた．5 次多項式が表現する典型的な曲線の全体的な形状を 11 タイプとし，その形状のパラメータ値のとり方により，さらに 13 種類の部分形状を導いた（図 2 参照）．

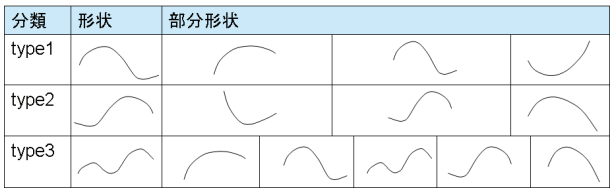

図 2: タイプごとに分類された部分形状 (一部)

図 2 に示す分類は，実際のコーパスから抽出されたグラフの挙動を説明するために使われる語彙表現の観点から導いた．グラフの全体形状を示す 11 のタイプはグラフのどの部分形状を含むかが決まっているため，５次多項式で認識されるグラフの形状は，始めに全体形状の特定のタイプを選別する．次に，その部分形状を数式的に解釈することにより最終的なグラフの形状を認識し，これを説明する適切な言語表現をする (図 3 参照)．

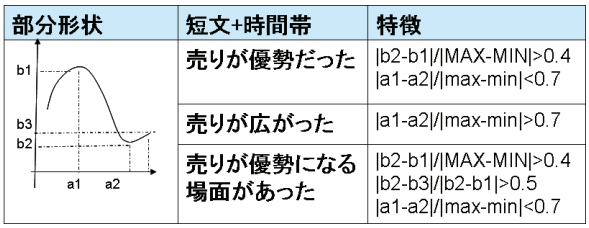

図 3: 部分形状の数式的解釈とその言語表現

2.3　辞書作成

辞書は，実際のコーパスとそれに対応する株価動向を示すグラフの部分形状の対応関係を観測することにより構築される（図 4 参照）．

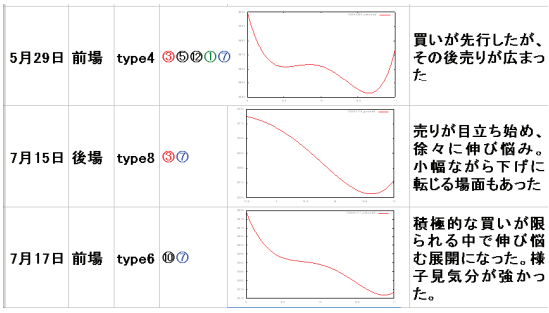

図 4: グラフの形状と言語表現の対応

辞書構築にあたっては，先行研究において使用された 2005 年 7 月 25 日から 8 月 30 日までの 27 個，2009 年 5 月 20 日から 7 月 24 日までの 20 個の株価動向を表す実際の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，文を収集し，辞書を構築した．構築された辞書は，図 3 に示すようにグラ

フの形状を数式的に解釈したものが語彙や文と対応するようにシステム内に実装されている．

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．

2.4 文法

テキスト②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．

2.5 実行例

システムは，3 つのウインドウで構成されており，MySQL に蓄積されている日経平均株価データを読み出し表示するウインドウ，グラフを表示するウインドウ，グラフの動向を言語で説明するウインドウからなる．図 5 では,「2009 年 10 月 14 日」と入力すると，以下のようなテキストが生成された．

「後場は，寄り付き後，売りが優勢だった．その後，
底堅さを確認した．その後，買いが入った.」

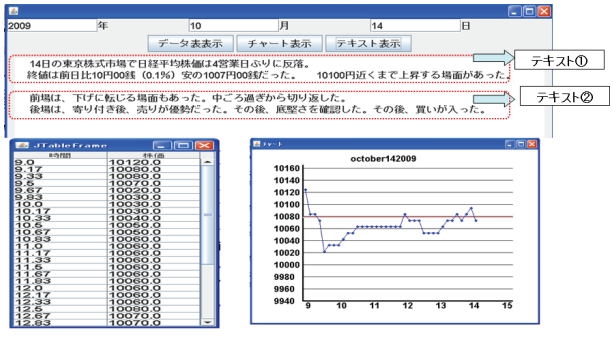

図 5: システムの実行例

3 システムの評価

表 1 に実際のコーパスとシステムによるテキスト生成結果を比較したものを示す．

表 1: 評価

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

辞書中の語彙表現を「状態」「変化率」「変動量」「その他」の 4 つの種類に分類し，それぞれに対して一致度を評価した (その他の項目には，“もみ合い” など方向性のないテキスト表現が入っている)．また，上図の「同意」を考慮した一致度とは，語彙が完全に一致していなくても,「売りが強まった」と「売りが広がった」など，同じグラフの挙動を意味しているものの一致度のことを指す.「同意」を考慮した一致度の定義としては，辞書のパラメータ設定から図 6 のように，同じ「売り」の挙動を示す語彙の中で包含関係となっているものを「同意」と考える．

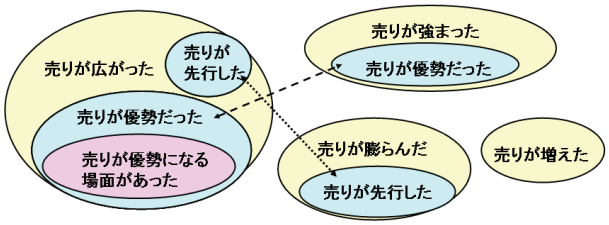

図 6: 言語表現の同意関係

表 1 より，同意を考慮した一致度の場合，システムのコーパスに対する一致度は 0.86(44/51) となり，作成した辞書が適切であり，また，言語化するグラフ挙動の箇所が同じである割合が高いといえる．一方，ひとつのグラフ挙動に対して，それが変化率の特徴で言語表現される場合もあれば，状態の特徴で言語表現される場合もあり，一概にグラフの同じ特徴に対して一致する生成されたテキストの数によってテキスト生成システムの性能評価はできない．このことを考慮して，表 1 中の列項目における「グラフの挙動表現に対する一致」は，同じ期間の同じグラフに対して生成されたテキストの内，グラフの各特徴において正しくグラフの挙動を表現するテキストの数を示している．この表の数値からは，実際のコーパスにおいて，対象とする期間内のグラフの挙動を 51 個（4+25+6+16）のテキストを用いて表しているのに対して，システムはグラフの挙動を表現する 47 個（11+12+11+13）の適切なテキストを生成していることがわかる．システムが生成したテキストの中には実際のコーパスに存在しないものもあるが，それはコーパスの作成者が気がつかなかったグラフの特徴をシステムが言葉で明示化したものとも考えることができる．このことからも我々の開発したシステムがグラフの挙動を説明するのに十分なテキスト生成が行えていることがわかる．

4 おわりに

本研究において，株価動向を示す時系列データの言語化手法およびその性能評価を示した．言語化手法において，辞書を強化することによりテキストの生成能力の向上を計り，また，性能評価についてはテキスト生成性能の評価を行うひとつの方法を提案した．今後の課題としては，より客観的な性能評価の指標としてグラフの形状認識において得点制を導入するなど，テキスト生成の性能評価指標を確立することや，株価のリアルタイムに基づく言語化などを行うつもりである．

参考文献

[1] 奥村菜穂子，グラフの挙動を表すテキスト生成，2005 年度お茶の水女子大学卒業論文，2005.

[2] 小林一郎，渡邊千明，奥村奈穂子：グラフとテキストの協調による知的な情報提示手法―日経平均株価テキストとグラフの提示を例にして，情報処理学会論文誌，48 (3)，pp.1058-1070,2007

[3] 加藤, 松下: 動向情報の要約・可視化から情報編纂へ, 第 21 回人工知能学会全国大会, 2H5-11, (2007).