解説

動向情報の要約と可視化†

— その研究課題とワークショップ —

加藤 恒昭*1・松下 光範*2・神門 典子*3

1. はじめに

動向情報とは時系列情報を総合的にまとめ上げることで得られるもので、製品価格や政党支持率の推移等がその例である。このような動向情報への関心に適切かつ効率的に応えるためには、様々な情報源から必要な情報を抽出し、簡潔で平易な文章や視覚的なグラフ、それらを協調させたマルチメディアプレゼンテーションとして要約・可視化して提示する必要がある。著者らは、この「動向情報の要約と可視化」研究の重要性と広がりに着目し、それに関連する研究を協調的競争的に展開することを目的としたワークショップを提案している。本稿では、まず、動向情報の要約と可視化について、その研究課題と位置づけについて解説する。続いて、著者らが提案するワークショップにおいて共通の素材として用いられる研究用データセットについて説明し、それをを用いることで、それらの課題への取り組みがどのような形となり、それらが如何に加速されるかを示す。

2. 処理の流れと技術

「動向情報の要約と可視化」技術は、動向に関する利用者の関心に、簡潔で平易な文章や視覚的なグラフ、それらを協調させたマルチメディアプレゼンテーションを用いて、適切に解答することを可能とするもので、「98年頃、パソコン業界はどんな感じでしたか」「去年から今年にかけて原油とガソリンの価格ってどう動いていますか」「去年の台風ってひどかったのだったけ」等の質問に答えるシステムの実現に繋がる。ここでは、

その要素技術について例を挙げながら検討し、課題解決の可能性を探っていく。まず、「98年頃のパソコン業界の動向」が利用者の関心であるとして、図1に示す処理の流れを追ってみる。

(1) **関心の理解・必要情報の決定**：利用者の関心を理解し、それに応えるために説明すべき情報（統計量等）を決定し、それらの関係を把握する。パソコン業界の動向が、パソコンの国内出荷台数、国内出荷額、それらのメーカー別シェアという統計量によって説明されることを得るとともに、それらの関係が把握される。これは知識工学指向の課題であり、オントロジ工学、セマンティック web 等の成果が期待される。なお、ある種の動向、去年の台風の動向等は、統計量の変化ではなく、「上陸」「発生」のような出来事に関する記述をまとめることで得られると考えられるが、その場合は、統計量の代わりにそれらの出来事が決定される。

図1 動向情報の要約と可視化の流れ

† Summarization and Visualizatoion on Trend Information
— Research Issues and Workshop —
Tsuneaki KATO, Mitsunori MATSUSHITA and Noriko KANDO

* 1 東京大学大学院 総合文化研究科 言語情報科学専攻
Language and Information Sciences, Graduate School of
Arts and Sciences, The University of Tokyo

* 2 日本電信電話(株) NTT コミュニケーション科学基礎研究所
NTT Communication Science Labs., NTT Corp.

* 3 国立情報学研究所ソフトウェア研究系
Software Research Division, National Institute of Informatics

- (2) **情報の収集**：新聞記事のような大規模テキスト集合や白書のような数値情報の集まりから、関心それ自体（パソコン業界の動向）や(1)で得られた統計量（国内出荷台数等）に関する情報を収集する。この部分については、現状の情報検索技術が充分その役割を果たすと考えられる。
- (3) **単位情報の抽出と分類**：収集された情報から必要な単位情報を抜き出す。テキスト情報の場合、検索される情報の単位は必ずしも必要な情報の単位と一致しない。例えば、検索の単位である記事から利用者の関心や必要な統計量に関連する文や節、例えば、パソコン業界の動向についての全般的なコメント「右肩上がりの急成長を続けてきたパソコン市場も、景気低迷によって踊り場にさしかかっている」⁹⁾や、それぞれの統計量に関する具体的な値の変化への言及「国内出荷台数は179万7000台で、対前年同期比9%増と1年3カ月(5四半期)ぶりに前年実績を上回った」等を抽出する。更にそれらの単位情報が何（統計量や出来事）についての情報であるかの分類を行う。
- (4) **情報抽出**：可視化もしくは要約のための情報をテキスト情報から抽出する。例えば、描画データとして利用可能な（統計量名、時点、値）の形式で、（国内出荷台数、1998年度第2四半期、179万7000台）等の情報を得る。
- (5) **要約**：個々の統計量や関心全体について文章による要約を作成する。(3)で抽出された単位情報を直接の入力とする場合と、(4)で抽出された情報から作成される場合とが考えられる。例えば「急成長を続けてきたパソコン市場は、98年前半には景気低迷により踊り場にさしかかったが、同年後半には低迷期を抜け出した」。なお、ここでの要約は必ずしも談話的整合性のある文章である必要はなく、その後の処理によっては動向の要点となる文の列挙で十分である。
- (6) **言語化**：数値情報を解釈することでその要約を得る。これについては本稿でこれ以上の議論をしないが、数値データを入力としてそこから傾向を言語もしくはシンボリックな形で得るための手法として、[3][20]等が提案されており、それと文章要約とを組み合わせると動向についてのレポートを生成する研究[1]も報告されている。また、単なる数値データではなくグラフを言語化する研究もある[4]。
- (7) **可視化**：(1)で得られた関心や(5)で得られた要約を考慮して、(4)で抽出された情報や数値データを用いて個々の統計量の可視化を行う。

- (8) **協調的統合**：(5)(6)で得られた要約や(7)で可視化された情報を、(1)で得られた利用者の関心や、統計量の間の関係、例えば「低価格化が進むと出荷台数の伸びほど出荷金額は伸びない」(このような関係は(3)で得られた情報からも付随的に獲得される)を考慮して、総合的にまとめあげ、利用者の関心に答えるプレゼンテーションを生成する。文章を中心にその中からグラフを参照するもの、グラフを中心にそこにコメントが註釈付けられているもの等、様々な形式が可能である。

ここでは、(4)―(8)の過程を Bottom-Up 的なものとして、文章による要約や個々の統計量に関するグラフが先に作られ、そこからプレゼンテーションが合成されるように示したが、Top-Down 的な過程、どのようなプレゼンテーションを作成するかの手組を先に決定して、その後に、そのために必要な情報を集め、その手組で要求される形式で可視化や要約を行っていくという方向も考えられる。

なお、以下の議論では、収集される情報をテキスト情報である新聞記事に限定し、そこからの要約や可視化を考えていく。この後の3節で、(3) **単位情報の抽出と分類**、(4) **情報抽出**、(5) **要約**、(7) **可視化**と(8) **協調的統合**についてその課題を検討していく。以下ではガソリン価格と原油価格を統計量の例とする。

3. 単位情報の抽出と分類における課題

検索されたテキスト情報から動向情報を得るのに必要な部分を抜き出すのがここでの処理である。このような単位情報の抜き出しは、個々の統計量に関する場合、「ガソリン」「価格」「円」等の単語を含んでいることである程度判断できる。しかし、その場合も以下に述べるような問題が残るうえ、利用者の関心そのものに関連する情報の抽出にはより洗練された処理が必要と考えられる。統計量の名称はその内部に意味的な関係を含んだ複合的な構成を持つので、その表現は様々であり、ある単位情報が実はどの統計量に言及しているかの判断は容易でない。以下の例文1-3は同じ統計量に関する記述であるがその表現はだいぶ異なる。省略や照応を考えると問題は更に複雑となる。これは情報抽出とも関連するが、例文4においても全国平均価格への言及があり、その値は91円である。

例文1 ガソリン価格（レギュラー1リットル）は全国平均で92円となり

例文2 レギュラーガソリン1リットル当たりの全国平均店頭価格は…92円となった

⁹⁾以下、引用はすべて1998年及び1999年の毎日新聞の記事から行っている

例文 3 全国平均のガソリン価格 (レギュラー) は…
1 リットル=90円から…

例文 4 全国 8 地区のうち東北、中部など 4 地区で、
ガソリン価格 (レギュラー、1 リットル) が前週
から 1 円上昇した。全国平均では 91 円で 変わらな
かったが

更に、着目している統計量に関する言及であっても、
事実報告である場合と予測や見込みである場合とがあ
り、これを区別する必要がある。動向情報については
前者のみを取り出すことが必要であり、例文 5, 6 等
は後続の情報抽出の対象とすべきでない。

例文 5 ドバイ原油が 1 バレル=20 ドルを超えれば、
1997 年 2 月上旬以来だ

例文 6 ドバイ原油価格は、現状の 1 バレル=20 ドル
を中心にした水準で堅調に推移するだろう

4. 情報抽出における課題

単位情報から、例えば、描画データとして利用可能
な (統計量名、時点、値) の形式で、情報を抽出する
処理である。統計量名は前述の処理で得られていると
して、時点と値の抽出が必要となる。時間表現の抽出
では、その時間表現が実際に「何時」を示しているか
という絶対表現を明らかにしなければならない [12]。
例えば「前週」「昨年 12 月」が実際に何時であるかがわ
からなければ、描画にも要約にも用いることはできな
い。値の抽出においても、例文 7 の「1 円」はガソリ
ン価格ではないので、ある助数詞を伴った数値表現を
単に抜き出せば足りるというものではない。

例文 7 昨年 6 月から続いていたレギュラーガソリン
の全国平均店頭価格の超安値 92 円が 1 円下がった

グラフ描画等を目的に複数文書からある統計量の値
の変化を抽出しようとした場合、ある出来事について
記述された文書からその出来事の情報のみを抽出する
場合に比べ、推論等の処理によってより多くの情報を
得られる可能性がある。上の例文 7 からは、推論や算
術演算によりこの記事の時点でのガソリン価格が 91 円
であることがわかるし、加えて、「昨年 6 月」からその
記事の直前まで 92 円という価格が続いていたこともわ
かる。例文 8 では、他の情報と合わせて初めて価格の
絶対値が得られる。また、最終的に時系列データとし
て並べた時に、「1 円」等の誤った抽出結果をはずれ値
として検出できる可能性もある。一方でそのまま抽出
したのでは描画に用いることができない情報も存在す

る。例文 9 は「～台」のような範囲表現を含むし、8
月のいつ 20 ドルを越えたのかも特定できない。このよ
うな情報は値に関する制約と捉えられるが、描画のデー
タとしてどう利用すべきかは難しい。この他にも「約
～」「～前後」等の蓋然的表現、統計量の単位や集約の
粒度のばらつき等、多くの問題が存在する [13]。

例文 8 ガソリン 2 円上昇—今月の全国平均、前月比
で

例文 9 原油価格は、… 8 月には、20 ドルの大台に乗
せ、先週末からは 22 ドル台での取引となっている

例文 10, 11 のような長期的な特徴の記述や定性的表現
をどう扱うかは大きな課題である。これらを要約に委
ねて情報抽出では扱わないという選択肢もあるが、定
性的な変化の記述を定量的な情報と重ね合わせる手法
が期待される。

例文 10 レギュラーガソリンの店頭小売価格 (…) は、
1992 年末の 125 円をピークに下がり始め、97 年 11 月
には 100 円台を割り込んだ

例文 11 原油の国際価格が昨年末から 50 % も急上昇し
ているにもかかわらず、末端のガソリン価格は逆
に緩やかな下落傾向にある

5. 要約における課題

文章による要約を生成する処理であるが、テキスト
情報 (抽出された単位情報) をその入力とする場合と
情報抽出によって得られた情報を入力とする場合とが
考えられる。

後者の場合、複数文書要約のひとつの枠組みとして
提案されている情報抽出と文生成を組み合わせる手法
[17] [23] が応用できる。これらの手法は、現在の情
報抽出技術が対象分野に依存する場合が多いのでその
制約を引き継ぎ、例えば地震に関する記事のみを要約
する枠組みとなっており、動向情報を対象として、ど
の程度の一般化が行えるかが課題となる。また、数値
情報の時間変化が重要となるが、数値情報の粒度を考
慮してまとめあげを行ったり、得られた数値情報の幅
を利用者に提示することが試みられており [23] [24]、
そのような数値情報の扱いに時間情報を組み合わせる
ことで、動向情報の要約に欠かせない値の変化の抽出
が行える可能性がある。

情報抽出を用いず、テキスト情報から直接要約を行
う場合、統計的な手法あるいはそれに統語的な情報を
加えて、類似した文を判定し、それらの頻度から重要
度を判定するとともに、重複した部分を抜き出すこと

で要約を作成する手法[2]の利用が考えられる。例えば、原油価格の場合、99年春から値上がりが続くが、それらに言及している記事において、例文12-14のように同年2月の底値が繰り返し言及される。このような節の類似性とそれら類似した内容の節の頻度に着目することで、値の変化の要所要所を取り出すことができる。一方、ある傾向、例えば値上がりが続いている間は、その度に新しい情報として異なる価格が記事中に現れるが、その場合でも、「急騰している」「原油続騰」「原油価格が急騰を続け」「原油価格の上昇が止まらない」「上昇が続いている原油価格も」の様に意味的に関連の強い表現や単語が繰り返し現れ、同一の傾向が続いていることが伺える。このような枠組みの中で、数値情報や時間情報をどのように扱い、取り入れていくかが課題となる。特にこれらは、その冗長性や矛盾が利用者にとって明白となるため、精密な扱いが必要となろう。

例文12 ドバイ原油価格は2月には10ドルだったので

例文13 ドバイ原油価格は今年2月には一時10ドルを割り込んでいたが

例文14 原油価格は、…2月の1バーレル＝10ドルから倍以上に高騰し

6. 可視化と協調的統合における課題

情報抽出等によって得られた情報を用いてグラフ等を作成する処理が可視化であり、得られたグラフや要約を有機的に協調させ統合する処理が協調的統合である。

一般に可視化という場合、多量のデータを捨象要約して視覚的な理解を容易にすることが多いが、テキスト情報から情報抽出を行った場合は、少量の不完全な情報を対象として推論等により情報を補完するような処理が必要となる。情報抽出との役割分担もあわせて検討していく必要がある。一方で、抽出されるデータ自体が動向の要点として特徴的であることも期待され、動向を概観するグラフが描きやすい可能性もある。

描画データだけで可視化が行えるわけではない。適切なグラフ形式を決定するためには、それによって何を示したいかという作成者の意図や関心、達成したい目的に注目する必要がある[5][8][18]。ガソリン価格については価格を縦軸、時間を横軸としたグラフが一般的であるが、原油価格のグラフとの統合にはいくつかの選択肢がある。「市場原理を失う石油業界－原油は上昇、なぜかガソリンは下落」が動向の要点であればそれが直観できるグラフの配置や組み合わせを考える必要がある。時間軸以外の軸が含まれる場合、例えば、

ビールの大手3社のシェアの推移では時間軸に加えて会社の軸を考える必要があるため、その幅が広がる。例えば、今までシェアのトップだったK社がその座をA社に譲ったグラフと、今まで業界の半分以上であったA社のシェアがそれを下回ったグラフは異なるものになると考えられる。動向を構成する単位情報が統計量ではなく出来事である場合は、更に様々な形式が可能になる。例えば、台風の上陸の場合、時間を軸にする必要は必ずしもなく、地図の上にそこに上陸した台風の情報をプロットしていく等が可能である。

グラフへの注釈付けは可視化のひとつの要点であり[14]、文章による要約とグラフとの協調的統合のひとつも捉えることができる。動向情報の場合、時間軸上でグラフと要約を関連づけられるので、「安定傾向」「急騰」「低迷」等の時点時点での評価をグラフに注釈づけることが可能となる等、注釈付けをはじめとして様々な協調的統合の可能性を持つ。

7. 研究の位置づけ

ここまで「動向情報の要約と可視化」における課題を具体的に眺めてきた。本節ではこれらの課題が様々な研究領域の今日的課題と密接に関連していることに触れておく。

動向に関する利用者の関心にマルチメディアプレゼンテーションを用いて適切に解答するための技術ということで、本技術は、まず、マルチメディアプレゼンテーション生成技術、マルチモーダル対話技術[6]である。この分野では一般にプレゼンテーション内容や対話内容の形式的な表現を前提としているのに対し、ここでは新聞記事等のテキスト情報を知識源に含めており、そのことで内容表現からの生成だけでなく内容表現への変換、つまり内容理解が合わせて必要となり、この点で新しい試みとなる。

質問応答技術、自動要約技術、情報抽出技術等の情報アクセス技術とも深く関連する。質問応答技術については、複数文書に散在した情報を収集しそれを要約して解答することが求められるという点で従来の質問応答の枠組み[7]を拡大し自動要約との融合を図るものとなる。自動要約技術の観点からは、動向情報の素材が複数の文書、例えば一年間に渡る新聞記事に散在しているため、複数文書の要約であり、しかもそれら全体の要約ではなく、利用者の関心に関わる部分のみを要約するという user-focused な要約である[15][16]。これらの点で、動向情報に関する研究は質問応答技術と自動要約技術の研究の最新の関心に合致している。情報抽出技術としては、動向情報という括りはあるものの様々な統計量に関する情報が抽出できる汎用的な

ものが必要である。汎用性・分野独立性の問題は、この技術の現在の課題でもある [19]。

動向情報の可視化、統計量の時間的な変化や複数の統計量の関係を直観的にわかりやすい形で視覚化するという課題は、動向情報という広い種類のデータに対する可視化技術の確立ということで可視化技術の共通基盤となる可能性を持つ。その一方で、可視化のデータをテキスト情報に求める場合、従来の多量のデータを捨象要約する方向とは逆に、少量の不完全な情報を対象として推論等により情報を補完するような処理が必要となり、そこに新しい可視化の枠組みが必要となる。また、情報可視化技術では膨大な情報を視覚的に整理し利用者が必要な情報に容易にアクセスできるようにするための研究 [11] [21] も大きな部分を占めており、それらは本技術の目的でもある。つまり、生成される要約や可視化をその背後にある文書等へのアクセス手段として利用するインフォメーション・アウトライニング [22] として本技術を捉えることができる。

この他、動向情報の要約と可視化は、言語理解技術やソフトコンピューティングにも関わり、言語表現と数値変化の関係について様々な研究素材を提供する。例えば、数値変化やその度合を表す程度表現にはどのような関係があるのか、数値変化と言語によるその捉え方やその伝え方とはどういう関係にあるのか等が重要となる。

8. ワークショップと研究用データセット

著者らは、「動向情報の要約と可視化」研究の重要性と広がりに着目し、それに関連する研究を協調的競争的に展開することを目的とした「MuST: 動向情報の要約と可視化に関するワークショップ」を提案している [9]^{§2}。このワークショップは、定量評価が可能な同一の課題に参加者が取り組むといういわゆる評価型のワークショップではなく、共通の素材を用いて、緩い意味で共通の課題に取り組むことによる議論と研究の活性化、コミュニティの形成や研究領域の認知度の向上、ツールやコーパス類の蓄積をその目的としたものである [10]。オーガナイザは、ML による議論の場の提供、研究成果発表の場の確保、研究の素材となる研究用データセットの提供を行っている。参加者には提供されたデータセットを用いた研究の成果や経過について指定された機会に発表することを求めている。

MuST の研究用データセットは、ワークショップの求心力となり、動向情報の要約と可視化に関する研究を加速するようにと設計された。このデータセットは、

この分野の様々な要素技術研究のハブとなるような、つまり、あるコンポーネントの出力であり別のコンポーネントの入力であるような処理の中間結果のコーパスや出力の参考となるような要約文章やグラフ等からなっている。

具体的には、情報源となるテキスト集合を、毎日新聞1998年、1999年の2年分とし、関心の対象となる動向情報として20のトピックとそれに関連する統計量や出来事のクラス（各トピックについて3つ前後、合計で58の統計量もしくは出来事）を選んで、それについての以下の情報を作成・収集した。

- それに関する情報を含んだ新聞記事のリスト (トピック単位)
- それら記事への注釈を行ったコーパス
- 記事中で述べられている統計量の変化や出来事に関する200字程度の文章による要約
- それらの統計量のその時期における変化や出来事に関する表やグラフ

選択したトピックとその統計量あるいは着目する出来事を図2に示す。トピックの後の括弧内に示したのは、収集された記事数である。既に述べたように、動向情報の素材は、大きく分けて、ある統計量のある時点（時間幅）での値に関する記述もしくはある出来事に関する記述のいずれかの形をしていると考えられる。ガソリン価格の場合は前者であり、個々の情報は、「ガソリン価格（レギュラー1リットル）は全国平均で92

ガソリン価格 (20)

レギュラーガソリン全国平均店頭価格、
ドバイ原油価格

ソニー（の連結業績） (9)

連結売上高、連結営業利益、連結経常利益、
連結最終利益

パソコン業界（出荷状況） (20)

国内出荷台数、国内出荷額、
メーカー毎の出荷台数シェア

ビール（および発泡酒）業界 (21)

出荷数量、メーカー別出荷数量、
メーカー毎の出荷量シェア、
ビール発泡酒総市場における発泡酒の割合

失業者 (14)

完全失業者数、事由別完全失業者数、完全失業率、
男女別完全失業率

台風 (15)

台風の上陸場所・日時（出来事）

大リーグのホームラン争い (13)

（マグワイヤ、ソーサ）のホームラン（出来事）、
（マグワイヤ、ソーサ）の年間ホームラン数

図2 トピック、統計量、出来事の例

^{§2}http://www.kecl.ntt.co.jp/sci/workshop/must/must_index.html

円となり、…」のような、ある時点でのある統計量の値の記述と、それと過去の値との比較や評価からなっている。一方、昨年の台風の動向の場合、中心となるのは「中型で並の強さの台風10号は、17日午後4時半ごろ鹿児島県枕崎市に上陸、…」のような個々の出来事の記述であり、その年にどこに幾つの台風が上陸したかという統計的な記述は少ない。本データセットではこれら両方の場合を扱っている。図中で「出来事」と注釈されているものが後者に分類される。

注釈の例を図3に示す。その仕様については文献[9]に詳しいが、単位情報の抽出結果に相当するものがunit要素として認定され、その分類として統計量(stat)や出来事(event)が属性として記述されている。これらのunit要素は事実報告だけでなく予測を内容とするものを含んでおり、予測の場合はtype="pros"が付与されている。unit要素の内部では、日付や統計量の名前や値等の部分が注釈づけられている。日付についてはその絶対表現(「昨日」に対して19990203等)と粒度を記述している。値は、統計量の値そのもの(val要素)と、その値の差や比、順位等の相対値を表現するもの(rel要素)とを区別している。更に、ひとつのunit要素のみで完結した情報の記述となるような補完(ins要素)を加えている。

この注釈付きコーパスは、前節の単位情報の抽出と分類の結果に相当するものであり、それに加えて、それに続く情報抽出において重要な時間表現の抽出とその絶対表現の算出の結果を含んでいる。従って、特定情報(ある統計量や関心そのもの)に関連する重要部分の抽出、時間表現の自動抽出およびその絶対表現の推定、事実情報と予測情報の分類等の研究に関心を持つ研究者にとっては、正解データあるいは学習用データとしての利用が可能である。そして、それらに続く処理である情報抽出・可視化・要約に関心を持つ研究

者にとっては、そこまでの処理に費す労力を節約して、関心のある部分に直接取り組むことを可能とする。例えば、可視化の研究を自然言語処理技術に関わることなく進めることができる。また、数値情報に関する様々な言語表現に関する研究等も、注釈情報を用いることで必要な分析を効率的に進めることができる。これらの研究者にとって、このコーパスは前処理が行われた入力データとなるし、予備分析済みの研究資料となる。動向情報の要約と可視化は、様々な研究領域にまたがった様々な要素技術を必要とし、そのシステム構築は必ずしも容易ではないが、このデータセットを用いることで、研究者は各人が関心のある要素技術に取り組むことが可能となる。求心力という言葉で表現したように、より広い視野に立った時に重要なことは、このデータセットが、今まで異なる分野に属すると考えられていた研究者達の議論を可能にするという点である。そして、同じ分野の研究者は、このデータセットを共通の素材とすることで一定の客観的評価が行えることになる。これらのことを通じた研究の加速と活性化が研究用データセット構築の目的であり、「MuST:動向情報の要約と可視化に関するワークショップ」の目指すところである。

9. おわりに

本稿では、動向情報への関心に適切かつ効率的に応えるために必要となる「動向情報の要約と可視化」について説明した。その研究は、質問応答、自動要約、情報抽出等の情報アクセス技術と研究的関心を共有するとともに、情報可視化についてもその基盤的技術や新しい枠組みを提示する可能性を持つ。加えて、この研究の共通の素材となる研究用データセットとそれを中心としたワークショップについてについて説明した。

この「MuST:動向情報の要約と可視化に関するワー

990617020 ガソリン価格

```
<unit stat="レギュラーガソリン全国平均店頭価格"><del> 石油情報センターが16日発表した給油所石油製品市況調査によると、</del><date gra="週" abs="19990617"> 今週</date> 調査の<name part="head"> ガソリン価格(レギュラー1リットル)</name>は<name part="foot"> 全国平均</name>で<val> 92円</val>となり、<date gra="週" abs="19990610"> 前週</date>の<name part="foot"> 平均</name>に比べ<rel> 1円</rel>上昇した</unit>。
<unit stat="レギュラーガソリン全国平均店頭価格"><name part="foot"> 全国平均</name>の上昇は<date gra="旬" abs="19990201"> 2月上旬</date>以来、<dur gra="月"> 4カ月</dur>ぶりだ</unit>。
```

981018004 台風の上陸

```
<unit event="台風上陸"> 中型で並の強さの<par> 台風10号</par>は、<date gra="時" abs="19980917"> 17日午後4時半ごろ</date><par> 鹿児島県枕崎市</par>に上陸、<date gra="時" abs="19980917"> 同日夜</date>には<par> 高知県宿毛市付近</par>に再上陸して進み<del>、西日本を縦断した</del></unit>。
```

図3 注釈された記事の例

クショップ」は、2004年11月にその提案を行い、2005年初めからMLの立ち上げ、参加者募集を行うとともに、データセットの一部の配付を開始し、同年4月より完全なデータセットの配布を行っている。NTCIR ワークショップ^{§3}のパイロットタスクとして位置付けられているため、毎日新聞記事の利用も可能である。ワークショップの研究成果発表の場としては、2006年2月から3月にクロズドのワークショップの開催を考えており、そこで参加者に何らかの成果発表、経過発表をしていただく予定である。加えて、ほぼ同じ時期にいくつかの学会の場を借りて、シンポジウムやワークショップを開催していく予定である。多数の方の参加をお願いしたい。

謝辞

本研究は、NTTと東京大学との産学連携共同研究、ならびに国立情報学研究所のNTTと東京大学との公募型共同研究によって支援されています。ご支援をここに感謝いたします。

参考文献

- [1] Saif Ahmad, Paulo C. F. de Oliveria, and Khurshid Ahmad. "Summarization of Multimodal Information" *Proceedings of LREC-2004*, 2004.
- [2] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. "Information fusion in the context of multi-document summarization" *Proceedings of ACL-1999*, 1999.
- [3] Sarah Boyd. "TREND: A System for Generating Intelligent Descriptions of Time Series Data" *Proceedings of the IEEE International Conference on Intelligent Processing Systems (ICIPS-1998)*, 1998.
- [4] Sandra Carberry, Stephane Elzer, et al. "Extending Document Summarization to Information Graphics" *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pp. 3-9, 2004.
- [5] Massimo Fasciano and Guy Lapalme. "Postgraphe: A system for the generation of statistical graphics and text" *Proceedings of 8th International Workshop on Natural Language Generation*, pp. 51-60, 1996.
- [6] 加藤 恒昭. "マルチメディアプレゼンテーションの自動生成に向けて - 自然言語生成からマルチメディア生成へ -" *情報処理*, Vol. 38, No. 12, pp. 1049-1056, 1997.
- [7] 加藤 恒昭. "自然言語による情報検索と類似性" *日本ファジィ学会誌*, Vol.13, No.5, pp. 436-444, 2001.
- [8] 加藤 恒昭, 松下 光範. "グラフを用いた探索的データ分析のためのマルチモーダル対話処理" *電子情報通信学会論文誌, D-II Vol. 187 No. 5*, pp. 1142-1152, 2004.
- [9] 加藤恒昭, 松下光範, 平尾努. "動向情報の要約と可視化に関するワークショップの提案" *情報処理学会研究報告*, 2004-NL-164, pp. 89-94, 2004.
- [10] 加藤恒昭, 松下光範, 平尾努, 神門典子. "評価なきワークショップの試み - MuST: 動向情報の要約と可視化に関するワークショップを例に -" *言語処理学会第11回年次大会併設ワークショップ「評価型ワークショップを考える」*, 2005.
- [11] Jock D. Mackinlay, George G. Robertson, and Stuart K. Card. "The Perspective Wall: Detail and Context Smoothly Integrated" *Proceedings of CHI'91*, pp. 173-180, 1991.
- [12] Inderjeet Mani and George Wilson. "Robust Temporal Processing of News" *Proceedings of ACL-2000*, pp. 69-76, 2000.
- [13] 松下 光範, 加藤 恒昭. "動向情報に基づく情報可視化の基礎検討" *人工知能学会全国大会*, 2005.
- [14] Vibhu O. Mittal. "Visual Prompts and Graphical Design: A Framework for Exploring the Design Space of 2-D Charts and Graphs" *Proceedings of AAAI-98*, pp.57-63, 1998.
- [15] 難波 英嗣, 奥村 学. "ここまで来たテキスト自動要約" *情報処理*, Vol. 43, No. 12, pp. 1287-1294, 2002.
- [16] 奥村 学. "テキスト自動要約" *情報処理*, Vol. 45, No. 6, pp. 574-579, 2004.
- [17] Dragomir R. Radev and Kathleen R. McKeown. "Generating natural language summaries from multiple on-line sources" *Computational Linguistics*, 24(3), pp. 469-500, 1998.
- [18] Steven F. Roth and Joe Mattis. "Data characterization for intelligent graphic presentation", *Proceedings of Conference on Human Factors in Computing Systems*, pp. 193-200, 1990.
- [19] 関根 聡. "情報抽出-情報を整理して提示する-" *情報処理*, Vol. 45, No. 6, pp. 563-568, 2004.
- [20] Somayajule G. Sripada, Ehud Reiter, et al. "Modeling the Task of Summarizing Time Series Data Using KA Techniques" In A. Macintosh, M. Moulton and A. Preece (ed) *"Applications and Innovations in Intelligent Systems IX, Proceedings of ES2001"*, 2001.
- [21] Russell Swan and James Allan. "Automatic Generation of Overview Timelines" *Proceedings of SIGIR-2000*, pp. 49-56, 2000.
- [22] 武田 浩一, 野美山 浩. "テキスト情報の可視化を利用した情報検索" *情報処理*, Vol. 41, No. 4, pp. 343-350, 2000.
- [23] Michael White, Tanya Korelsky, et al. "Multidocument Summarization via Information Extraction" *Proceedings of HLT-2001*, 2001.
- [24] Michael White, Claire Cardie, et al. "Detecting Discrepancies in Numeric Estimates Using Multidocument Hypertext Summaries" *Proceedings of HLT-2002*, 2002.

(2005年5月6日 受付)

^{§3}<http://research.nii.ac.jp/ntcir/>

[問い合わせ先]

〒153-8902 東京都目黒区駒場 3-8-1

東京大学大学院 総合文化研究科 言語情報科学専攻

加藤 恒昭

TEL: 03-5454-6302, FAX: 03-5454-4329

E-mail: kato@boz.c.u-tokyo.ac.jp

著者紹介

かとう つねあき
加藤 恒昭 [非会員]

1959年生まれ。1981年東京工業大学工学部電気電子工学科卒業。1983年東京工業大学大学院総合理工学研究科電子システム専攻修士課程修了。同年、日本電信電話公社（現 NTT）に入社。2000年より、東京大学大学院総合文化研究科言語情報科学専攻助教授。現在に至る。質問応答、対話処理、マルチモーダルコミュニケーションに関する研究に従事。工学博士。電子情報通信学会、人工知能学会、言語処理学会、ACL 各会員。

かんど のりこ
神門 典子 [非会員]

1994年慶應義塾大学文学研究科博士課程修了。博士（図書館・情報学）。同年学術情報センター助手。1995年米国シラキウス大学客員研究員。1996～1997年デンマーク王立図書館情報大学客員研究員。1998年学術情報センター助教授。2004年より国立情報学研究所教授。現在に至る。テキスト構造を用いた検索と情報活用支援、情報アクセスシステムの評価等の研究に従事。ACM-SIGIR, ASIS&T, 言語処理学会、人工知能学会、情報処理学会、日本図書館情報学会各会員。

まつした みつのり
松下 光範 [正会員]

1969年生まれ。1993年大阪大学工学部精密工学科卒業。1995年大阪大学大学院基礎工学研究科物理系専攻制御工学分野博士前期課程修了。同年4月、日本電信電話㈱入社。現在に至る。情報可視化、自然言語理解、インタラクションデザインに関する研究に従事。博士（工学）。平成14年度情報処理学会論文賞等受賞。情報処理学会、人工知能学会、ACM 各会員。