新聞記事におけるテキストとグラフの協調に関する分析

加藤 恒昭
kato@boz.c.u-tokyo.ac.jp
東京大学

松下 光範
mat@cslab.kecl.ntt.co.jp
日本電信電話株式会社

1．はじめに

テキストと統計グラフとを協調させて質問に回答するマルチモーダル対話システムについて研究を進めている．このようなシステムにおいては，回答に含めるべき情報をテキストと統計グラフとでどのように役割分担して伝達するかの方針が重要となる．その予備検討のために，ひとかたまりの事実を伝達している既存のメディアとして，統計グラフを含み事実を報告している新聞記事をとりあげ，その見出しとなっているテキストとグラフとの間での情報の重なりや関係を分析した．

2． 分析

2．1 分析対象

一般全国紙3紙（朝日新聞，毎日新聞，読売新聞）の朝夕刊各3ヶ月分（99年5月，02年4月，7月）について，そこに現れる記事のうち，以下の条件を満たすものを分析対象とした(分析には縮刷版を用いた)．これらは，マルチモーダル対話システムが質問に対して事実を簡潔に回答するという状況の参考になるものに分析対象を限定するための条件である．

- 棒グラフ，折線グラフ等の統計グラフを含む
- 定期的に掲載されタイトルが付けられているコラムではなく，事実情報を伝えるいわゆる報道記事である
- 紙面の半分以上を占めるような特集記事（例えば地方統一選挙の結果の分析）ではない
- それ以前の面に関連記事を持たない

調査した延べ9ヶ月で，この条件にあう記事は265件であった．これらの記事の殆ど，254件はグラフを一枚だけ含んでいる．

2．2 分析方法

コーディングは以下の様に行った．(1) 見出しを，活字の種類や記事中での配置を考慮し，ひとつひとつの事実や主張などを単位として分割する．見出しには時期や組織名などを示す名詞も多いが，それらも一つの単位とする．以下，これら一つ一つを見出しと呼ぶ．(2) 見出しの中で最も大きな活字で述べられているものをその記事が伝える主たる情報と考え，主見出しとして選択する．(3) 各見出しをその意味内容に基づいて分類する．(4) 見出し間の修辞関係を分類する．

意味内容に基づいた分類(3)では，事実や主張である見出しについて， 程度や量や順位の比較や変化について述べており，具体的な数値を含むもの（PN1），同じく程度や量について述べているが，具体的な数値を含まないもの（PN2），これら以外で事件や出来事が起こったことを述べているもの（PE），その他，状況報告，評価，判断，見解，第三者発言の引用等（PO）の4種類に，それ以外の見出しは，時期や数値（TN）と，組織名称等それ以外のもの（TO）の2種類に分類した．ここで，事実や主張であるかは，統語的な特徴よりも内容に基づいて判断している．例えば，「市場に漂う楽観ムード」「市場は『一時的』の見方」等はTOではなく，POに分類している．

修辞関係の分類(4)で用いた修辞関係は，具体化や例示等，関係先の内容をより詳細に説明している（詳細化），関係先の内容とほぼ並んで述べられるような内容である（並列），関係先の内容をより詳細なものとする内容を加えている（補足），ひとつの事実や主張の主部と述部が別々の見出しとなっている場合に述部となっている関係先の主部である（主題），調査の方法，対象，年度等，関係先の内容の情報源や条件を示している（条件），関係先の内容が表わす状況や出来事の原因・背景・根拠・理由を示している（原因），関係先が表わす状況や出来事の結果や影響として生じた事態や状況を示している（結果），関係先が表わす状況に対するコメントや評価，もしくはその状況への対策の提言である（評価）の7種類である．

付録に分析例を示す．

これらの分析を施した見出しに対して，その内容とグラフによって伝達される情報との関係を明らかにす

るために，以下の分析を行なった．まず，見出しの内容がグラフが読み取れるかを調査し，(A) 読み取れる，(B) 見出しの前提部分が読み取れる，(C) 直接は読み取れないが強い関連がある，(D) 読み取れない，に分類した．ここで，「(A) 読み取れる」の定義として，「急増」等の定性表現 (PN2) については，その感触がグラフからつかめればよいとし，「3年ぶり」「3000万を越す」等，具体的数値がある場合 (PN1) は，それが読み取れることとした．(B)は「加入者減少に悩む」に対して加入者減少を示すグラフがあるような場合である．なお，他と「条件」や「主題」の関係にある見出しは，この分析の対象外とし，他と「補足」の関係にあるものは，関係先と合わせて考える，つまり，「東証終値1万7000円台回復」と「1年2ヶ月ぶり」の場合，「東証終値1万7000円台回復」が読み取れれば，前者に(A)，更に「東証終値1年2ヶ月ぶりに1万7000円台回復」が読み取れれば，後者にも(A)を付与することとした．

更に，グラフに対して，見出しで伝達されているもの以外でどのような情報を伝達しているかを分析した．この分析についての詳細は次節で結果と共に述べる．

3 結果

3．1 見出しとグラフで伝達される情報の重なり

複数のモードで伝達される情報にどの程度の重なりがあるかについては相反する二つの予測が可能である．

- 言語モードと視覚モードの役割分担がなされているはずなので，見出しとグラフとの情報の重なりは比較的少ない．
- 見出しとグラフは，共にその記事で最も重要な情報を伝えているはずなので，情報の重なりは極めて多い．

この点に関する結果を図1に示す．主見出しの内容についてみると，97件 (37%) がそれをグラフから読み取ることができた．このうち，38件 (全体の14%) については，主見出しを含むすべての見出しの内容をグラフが読み取ることができた．見出しのいずれかの内容がグラフから読み取れるものは151件 (57%) で，逆に53件 (20%) の記事ではいずれの見出しの内容も読み取ることができなかった．付録に典型的な例を幾つか示す．

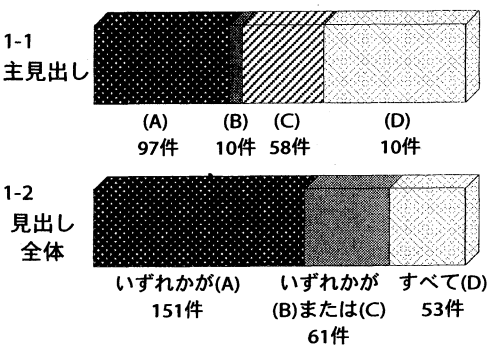

図1 見出しの内容がグラフから読みとれるか

見出し毎に見てみると，数量等に関連する内容を持っているということで，PN1もしくはPN2に分類された303見出しのうち，173見出し (57%) の内容が(A)グラフから読み取ることができ，47見出し (16%) が(C)グラフと強い関連を持っていた．

3．2 見出しとグラフの関係

86見出し (そのうち主見出しは58) が(C)グラフと強く関連していると分類されたが，その関連の仕方は以下の6つに分類することができた．

C1 グラフに描かれている統計量 (従属変数) は見出しの内容と強く関連するが，厳密には同じものではない．例えば，見出し「受験者数増大」対して受験倍率の変化を示すグラフ．

C2 グラフに描かれている期間等，独立変数の範囲が不適切である．例えば，見出し「10年ぶり」に対して5年分のデータを示したグラフ．

C3 見出しで述べられている状況を表す統計量を描いたグラフであろうことは推測できるが，ある程度の知識を必要とする．例えば，見出し「事業が好調」に対して販売量と生産量の増加を示すグラフ．「美術収集熱高まり」に対して美術品オークション年間落札価格の増加を示すグラフ．

C4 見出しの表現が比喩的，慣用的で，グラフの内容と直接結びつくかの判断が難しい．見出し「バブル弾ける」に対して株価急落のグラフ．見出し「○○大人気」に対してその製品の売上高の急増を示すグラフ．

C5 見出しの内容の具体例を示すグラフであり，見出しの内容の全体を表現していない．例えば，見出し「アジアでインフレ」に対してシンガポールのイ

ンフレを示すグラフ．見出し「ネット関連株一斉に急落」に対してヤフーの株価を示すグラフ．

C6　その他，これらのいずれにも分類できない．

ここで，C1,C2は，見出しとグラフに若干のズレがあるような状況であり，C3-C5は，グラフが見出しの具体化や例示等の詳細化となっている状況である．今回の調査での出現数を図２に示す．見出しの内容を直接表現していない場合も，具体化や例示として見出しと関係するものが多いことがわかる．

３．３　グラフで伝達される情報の位置付け

主見出しの内容がグラフから読み取れず（(A)以外），主見出し以外の見出しの内容がグラフから読み取れる場合に，その見出しが主見出しとどのような修辞関係にあるかを分類した．結果を図3に示す．図3-2は，比較のためにすべての見出しの修辞関係の出現比率を示したものである．図3-1からは，主見出しに対して原因（背景・根拠・理由）の関係を持つ見出しの内容がグラフとして表現されることが多く，次に詳細化（具体化，例示）の関係にあるものが続いているのがわかる．ただし．補足，主題，条件，評価の関係にある見出しの内容はグラフとして表現されえない，もしくは極めてされにくいことを考慮すると，図3-1と図3-2での修辞関係の出現比率に大きな違いはなく，結果の解釈には慎重であるべきである．

また，図１に示したように，主見出しの(B)前提部分が読み取れるものが10件（4%）あるが，この前提部分が別の見出しであれば，主見出しとの関係は原因となろうから，この場合も主見出しに対して原因の関係にある見出しの内容がグラフとして表現されていると考えることができる．

３．４　グラフのみで伝達される情報

これまで，見出しで伝達される内容がどれだけグラフでも伝達されているかをみてきた．一方で，見出しでは伝達されずグラフのみで伝達されている情報もある．これを分析する場合，グラフで伝達される情報の記述が難しいことが問題となる．例えば，付録のグラフは，その時点までの各月の支持率不支持率を示しているので，見出しに対して格段に情報が多いと考えることもできれば，このグラフは２つの見出しが伝える「逆転」「まで低下」の情報を伝えるためのもので，見出し以上に情報は伝えていない（グラフでこれらの情報を伝えようとすると，このようなグラフにならざるをえない）と考えることもできる．今回の調査では，比較的客観的に判断できる以下の特徴を取り上げ，集計した．

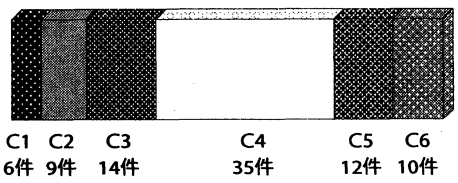

図2　見出しとグラフとの関係

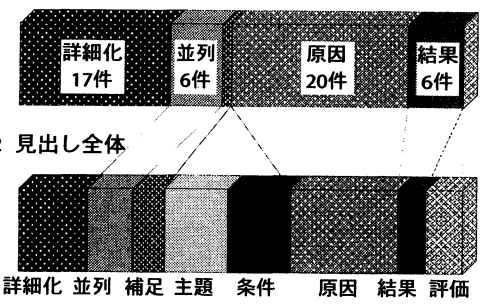

図3　グラフで伝達される情報の位置づけ

AS　見出しで言及されていない統計量についてのグラフである．もしくは，見出しで言及されている統計量に重ね合わせて別の統計量も示されている．例えば，経常利益について述べている見出しに対して経常利益だけでなく売上高が描かれている．

DF　特に円グラフや帯グラフにおいて，対象（独立変数）の分類が細かい．例えば，見出しでは，子供，成人，老人で述べられているのに，10歳毎の分類のグラフとなっている．

TL　時間を軸としたグラフにおいて，描かれている期間が必要以上に長い，あるいは現状のみに言及した見出しに対し過去のデータも描かれている．

CO　比較のためか，同じ統計量について，他の対象の値も描かれている．例えば，日本のコンピュータの普及率に関する見出しに対して米国の普及率も描かれている．

これら特徴を持ったグラフの数を図４に示す．なお，AS96件のうち53件は，見出しでは言及されていない統計量だけについてのグラフである．

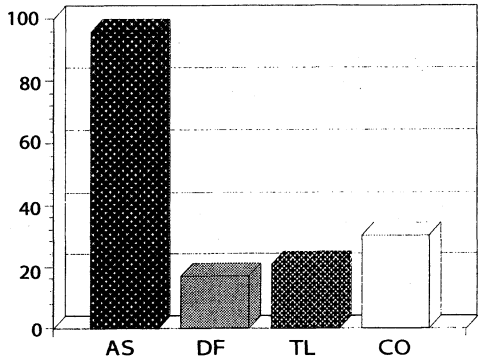

図4　グラフのみで伝達される情報

4.　おわりに

統計グラフを含み事実を報告している新聞記事をとりあげ,その見出しテキストとグラフとの協調について分析した.両モードで伝達される情報の重なりは比較的大きく,分担して情報を伝達するというより主たる情報を複数のモードで重複して伝達する傾向にある．両モードで伝達される情報が同じでない場合には，グラフが具体的なデータを示し，テキストはそれを抽象的な表現でまとめている.主たる情報がテキストで伝達され,それを中心もしくはまとめとする状況全体がグラフで伝達されるという構図が感じられる.この点では，情報の重複を恐れず，重要な情報は複数メディアで伝達すべきという方向が見える．一方で，主たる情報が表現する状況や出来事の原因となる背景をグラフが伝えることも多いが，これはデータを具体的に伝えるというグラフの特徴からくる役割分担とも考えられる.

3.4でも述べた様に，テキストとグラフとの協調をより精密に考察するためには,あるグラフがどのような情報を伝達しているかの分析が重要となる.この分析については，ある情報（だけ）を伝達するためにどのようなグラフを描くべきかという設計と相対であるので,マルチモーダル対話システムの設計と合わせて考えていきたい.

付録　分析例