新聞記事の数値による情報検索システムの提案と実装

杉浦隆博[††] 吉田稔[†] 山田剛一[††]
増田英孝[††] 中川裕志[†]

Trend Information Extraction with Numeral Information from Newspaper Articles

Takahiro Sugiura,[††] Minoru Yosida,[†] Kouichi Yamada,[††]
Hidetaka Masuda[††] and Hiroshi Nakagawa[†]

1. はじめに

計算機の処理能力の向上，またネットワーク環境の普及に伴い，ユーザが利用可能な情報は増化の一途を辿っている．これに伴いユーザの関心や興味と合致する情報を，より直観的かつ簡易に提示するための技術が求められている．これらの要求に応える技術に動向情報を対象としたものがある．

動向情報とは，商品の価格や売上高，内閣支持率などのように，時系列変化に伴って変動する統計量のことである．このような動向情報は単なる一次元の時系列情報ではなく，製品のシェアのように複数の企業が関係したり，地域毎の土地価格の変動のように空間的な広がりを持ったりするなど，複数の主体や空間軸などが関係する多次元情報である．

近年，これらの動向情報を対象とした複数文書要約や可視化に関する研究が活発化している 1)2)．しかし，動向情報を新聞記事などの文書群から自動的に抽出するのは現在でも困難であり，いくつかの先行研究は存在するが，要約や可視化などを行う際には，人手でタグ付けを行ったコーパスを使用する場合が多数である．

また，数量表現抽出の研究では，係り受け構造と優先規則による抽出規則に基づく抽出方法 3) が存在し，数量表現と対応する事柄の抽出に関して高い再現率と適合率を得ている．

しかし，新聞記事などの文章中に出現する数値には，同一文書内の複数の数値情報の関係に意味を持つものが存在する．例えば，ビールメーカの出荷シェアに関する数値情報には，「キリンが 42.6%」,「アサヒが 32.7%」といったものが存在するが，これらの数値情報は「ビールメーカの出荷シェアはキリンが 42.6%でアサヒが 32.7%である」といった形で表せるとき初めて有効な数値情報であると言える．

そこで本研究では，同一記事中に存在する「42.7%」や「32.7%」といった数値情報が「ビールメーカ各社の出荷数量」というトピックで結び付き、それぞれの数値が「アサヒ」、「キリン」といったビールメーカの出荷数量である、といった数値情報の関係性を抽出することを目的とする。

そして，本稿では複数の数値情報を関連づけるために必要な情報として，特に「相対値」に着目し，統計量値の候補となる数値情報をその統計量名、統計量の相対値と合わせて抽出を行う。また、抽出した数値情報を用いた記事の検索、提示、並び替えなどの機能を提供するシステムを提案する．

本稿の構成は以下のとおりである．次節では，数量表現抽出に関連する研究について述べ、3 節では既存の動向情報コーパスについて述べる。4 節では，新聞記事からの動向情報の自動抽出に関する手法と手順について説明する．本研究では，抽出した動向情報を用いて情報検索システムを実装したが，4 節では，その情報検索システムについて述べる．5 節では，本稿のまとめ，及び今後の課題について述べる．

2. 既存の動向情報コーパス

現在，動向情報に関する研究では，人手でタグ付けを行ったコーパスを使用する場合が多数である．これらのコーパスは，記事中の株価や商品の出荷数量などの統計量に対し，要約や可視化に必要となるタグ付けを行っている．本稿では「動向情報の要約と可視化に関するワークショップ (略称 MuST) における研究用データセット 4)5)」にある動向情報コーパスを例として取り上げる．

このコーパスは，各記事に対して，統計量の名前や値，日付などの要素を抜き出し，値に関してはどの統計量のものか，日付に関してはその絶対表現はいつかを記述したものである．

以下は毎日新聞の PC 出荷シェアの記事にタグを付与したものである．

```
<unit stat="メーカー毎の PC 出荷シェア">
    <par> NEC など昨年の上位 5 社 </par>
    の
    <name> シェア </name> は
    は
    <pro ref="前年比" id="9801220"> 同 </pro>
    <rel type="prop">3.1 ポイント </name >
    低い
    <val>82.7%</val>
    となった
</unit>.
```

タグの詳細な使用に関してはを表 1 に記す．

本稿では，動向情報の中で最も重要な情報となる，統計量の名前，値，値の相対表現の自動抽出を行う．

† 東京大学情報基盤センター図書館電子化研究部門
Language Infomation Laboratory, Information Technology Center, Tokyo University
†† 東京電機大学工学研究科情報メディア専攻
Graduate School Engineering, Tokyo Denki University

表 1 コーパスで使用するタグの意味

タグ	意味
<unit>	動向情報の統計量や出来事に言及している部分を示す．言及している統計量 (stat) や出来事 (event) が属性として付与されている．
<name>	統計量の名前を示す．
<date>	動向情報に関する時刻を示す．
<val>	統計量の値を示す．
<rel>	統計量の値そのものではないが，その値の差や順位，比などの相対値を示す．
<pro>	参照表現を示す．

3. 数値情報の自動抽出

本研究では，毎日新聞 98 年版 6) と毎日新聞 99 年版 7) の新聞記事を対象に数値情報の抽出を行う．本稿では，例として以下の記事から動向情報の抽出を行う．

> シャープが２９日発表した０２年９月連結中間決算は，液晶カラーテレビやカメラ付き携帯電話などの戦略商品の好調で売上高が前年同期比７．８％増の９７１７億円，経常利益が同２１．３％増の３８３億円，最終（当期）利益が同４０．５％増の２２８億円と大幅な増収増益になった．

この文章はシャープの 9 月の連結中間決算に関する記事である．統計量として，売上高や経常利益，そして最終利益が記事中に含まれている．本節では，これらの統計量の抽出手順について述べる．

3.1 係り受け解析

まず，新聞記事に対し係り受け解析を行う．係り受け解析には，日本語係り受け解析器 Cabocha を使用する．次に，係り受け解析を行った新聞記事の各文節を文末から辿り，依存構造木の形に変換する．依存構造木に変換したの結果が図 1 になる．依存構造器の要素間の関係，つまり記事中の文節間の関係を基に動向情報の自動抽出を進める．

3.2 数値情報の抽出

新聞記事の依存構造木への変換を行った後に，数値情報の抽出を行う．ここでいう数値情報とは統計量の値の候補となるものである．数値情報の特定には，Cabocha による係り受け解析を行う際に取得する品詞情報を利用する．品詞情報が「名詞−数」である形態素を持つ文節を，数値情報とする．また、数値の単位に関しては、品詞情報が「 名詞−接尾−助数詞」となる形態素の抽出を行う。

3.3 統計量名の推測

統計量名の候補として，3.2 で抽出した数値情報の要素と関係する以下の要素を取り出す．

(1) 数値情報の子要素 (数値を修飾する文節)
(2) 数値情報の親要素 (数値の係り先の文節)
(3) (2) の要素の数値情報以外の子要素

数値情報「9717 億円」に対して,「7.8%増の」,「なった」,「経常利益」,「売上高」などが候補になる．このとき, 数値「9717 億円」と関係を持つ動詞は「なった」となり, 数値が目的格だと判別できる．そのため, 数値「9717 億円」を表す統計量名は主格となる必要があると仮定する．主格は目的格である数値より前に出現するものであるとして抽出し, ここでは「売上高」が統計量名の候補となる．

表 2 コーパスで使用するタグの意味

単位	対応する表現
円	売上高，株価，利益，費用，利子，負債 etc
人	人口，客員，動員，生徒数

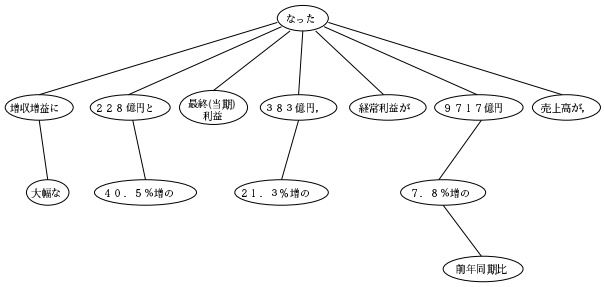

図 1 依存構造木

3.4 統計量名の特定

3.2 で選出した候補から, 統計量名を特定する．統計量名の特定には, 数値情報が持つ単位と対応する表現を基準にして行う．ここでは, 数値情報「9717 億円」の単位は「円」であるため, 統計量名として「株価」,「利益」,「売上高」といった表現を含む文節が統計量名となる．単位と対応する表現の組合せを表 3 で示す．3.2 で選出した候補が「売上高」であるため, 数値情報「9717 億円」の統計量名は売上高となる．

3.5 統計量の相対値の特定

3.1 における依存構造木の要素を対象とし, 統計量名の相対値を特定する．ここでは, 文節中に「%」,「割」などの比率を表す語を含むものを統計量名の相対値とする．

4. 抽出結果の評価

4.1 再現率の評価

再現率に関しては，MuST で配布している動向情報コーパスを正解データとし，1998 年から 1999 年までの毎日新聞記事中に出現する MuST コーパス中にある統計量の値と組となる，統計量名，統計量名のパラメータ，そして統計量の相対値が正しく抽出できたものを正解とする．

MuST コーパス中にある数値情報に関連する抽出結果に対して，MuST コーパスと同様の統計量名，統計量の相対値が完全に抽出できた場合のみ○，不完全に抽出したものを△，抽出できたんかったものを×とすることで再現率を評価する．統計量名と統計量の相対値のそれぞれを,「物価」や「内閣支持率」といった MuST コーパスの各トピック毎に評価している．

図 2 と図 3 が、各トピックの再現率に関する評価結果であり，表 3 が抽出結果全体の再現率の評価結果である．

評価結果から，統計量名の再現率に関しては「ソニー」,「エアコン」,「商業販売統計」といったトピックが比較的高く，反

対に「ガソリン」,「長野五輪」,「物価」といったトピックは低い再現率となった.

また，相対値の再現率に関しては「ソニー」,「商業販売統計」,「百貨店」とったトピックが高く，反対に「ガソリン」,「住宅」,「景気予測」,「総合電機 3 社」といったトピックは低い再現率になっている.

表 3 再現率の評価

	統計量名	統計量の相対値
再現率	35.07%	28.85%
再現率 (△を含む)	42.34%	29.15%

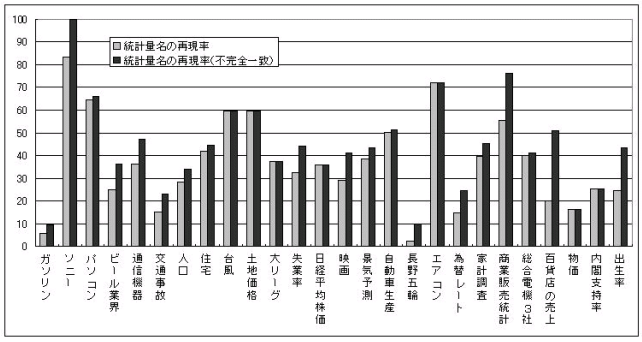

図 2 統計量名の再現率

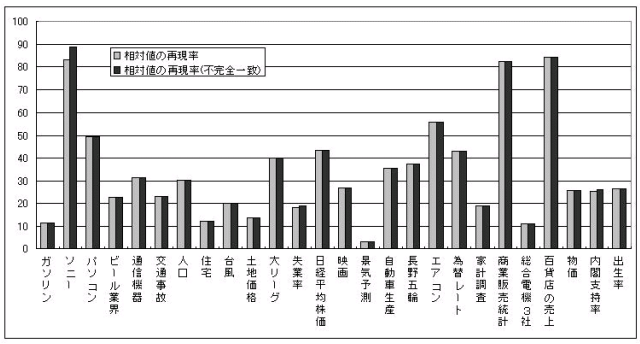

図 3 統計量の相対値の再現率

4.2 適合率の評価

適合率は，MuST コーパスに存在するトピックと関連する 1998 年から 1999 年の毎日新聞記事を対象とし，対象記事中に存在する統計量の値と組となる，統計量名，統計量の相対値が正しく抽出できたものを正解としている．MuST コーパスでは取り扱っていない数値情報 (例：ビールメーカの各製品の出荷数量等) も評価対象とする。そのため、正解の判定は人手で行い、数値情報に関係する統計量名，統計量の相対値に対して評価する．適合率の評価は，再現率と同様に各トピック毎に評価している.

図 4 と図 5, 及び表 4 が適合率に関する評価結果となっている.

評価結果から，統計量名の適合率に関しては「ソニー」,「商業販売統計」,「百貨店の売上高」といったトピックが比較的高く，反対に「ガソリン」,「人口」,「長野五輪」,「花粉」,「為替レート」といったトピックは低い適合率となった.

また，相対値の適合率に関しては「ソニー」,「百貨店」とったトピックが比較的に高いものの，他のトピックに関しては低い適合率となった.

表 4 適合率の評価

	統計量名	統計量の相対値
適合率	57.09%	37.84%

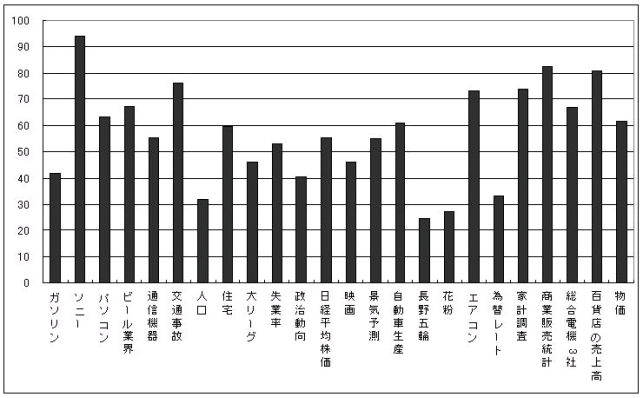

図 4 統計量名の適合率

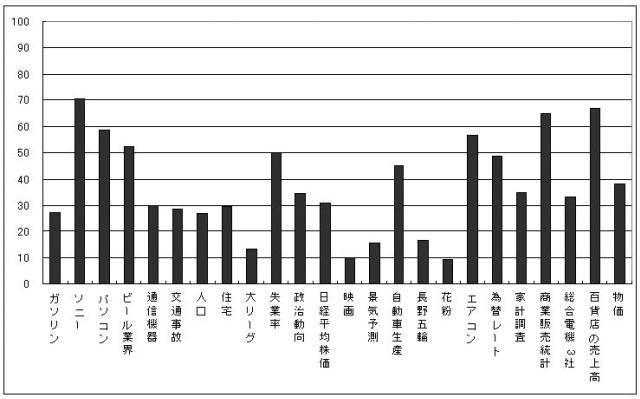

図 5 統計量の相対値の適合率

5. 評価結果の考察

5.1 統計量名の抽出結果に関する考察

再現率の評価と適合率の評価の両方で，比較的良い評価結果が得られた「ソニー」,「商業販売統計」,「エアコン」といったトピックは，統計量の値に対応する統計量名が「売上高」,「経常利益」,「出荷台数」,「販売額」など出現傾向が明らかであった．そのため，単位と対応する統計量名を決定しやすく，良い再現率と適合率を得ることができたと言える.

これに対し,「長野五輪」,「為替レート」,「人口」といったトピックは再現率と適合率が低くなっている.

評価結果が低かった理由はそれぞれのトピック毎に異って

おり,「人口」に関しては以下のような文構造において，統計量名を確定することができなかったためである．

> ６５歳以上は１９７９年に１０３１万人と１０００万人を上回り、１２年後の９１年には１５５８万人と１５００万人を超えた．

上記の文構造の場合,「１５５８万人」と「１５００万人」という数値情報に関しては正しい統計量名のパラメータである「６５歳以上 (の人口)」が抽出できるが,「１０３１万人」と「１０００万人」という数値情報に関しては，係り受け構造上，本手法では抽出することが困難であり，再現率と適合率が低下してしまった．

「長野五輪」に関しては，以下のような表形式の文構造において，数値情報に関係する統計量名を抽出することが不可能であるためである．

◇国別獲得メダル表◇

	金	銀	銅	計
ノルウェー	５	６	３	１４
ドイツ	５	４	４	１３
ロシア	５	３	１	９
日本	２	１	１	４
……				
	計２９	２９	２９	８７

また，上記の表形式の文構造は「長野五輪」以外のトピックでも出現し，その場合でも数値と関係する統計量名を抽出することは不可能である．

そして,「為替レート」に関しては「１ドル＝１３６円台」といった，文の一文節中に数値情報とそれに対応する統計量名が出現しており，現在の抽出手法ではこのような場合を考慮していないため，評価結果に影響を与えている．

5.2 統計量の相対値の抽出結果に関する考察

統計量の相対値の抽出結果は，再現率と適合率の両方において，統計量名の抽出結果よりも低い値となっている．これは，統計量名の相対値が統計量名とは異り，必ず数値情報と組になるものではないからである．現段階では，数値情報と関係する相対値の有無の判定を行っていないため，統計量名よりも低い評価結果となっている．また，統計量の相対値となるものが「前期比４５・２%増」といった比率による表現や,「過去最低」「最安値」といった順位による表現といった比較的に相対値であると判定しやい表現以外の,「人口増可は３０万人」や「１５００円の減少」といった表現を相対値と特定するのが困難であることも一因である．

6. 数値による情報検索システム

本研究では、抽出した統計量の値と統計量名，統計量の相対値を利用して、新聞記事上の数値による情報検索システムを実装した。このシステムは、検索キーワードに関係する数値情報を利用者に提示するシステムである。検索対象の新聞記事は、数値情報の自動抽出を行った毎日新聞 98 年版 6) と毎日新聞 99 年版 7) である。動向情報検索システムの検索結果画面が図 6 である。本システムでは，検索結果を数値によって並び替えることがが可能であり，売上高に関して検索を行えば，新聞記事中に存在する売上高の順に検索結果を提示することが可能である．

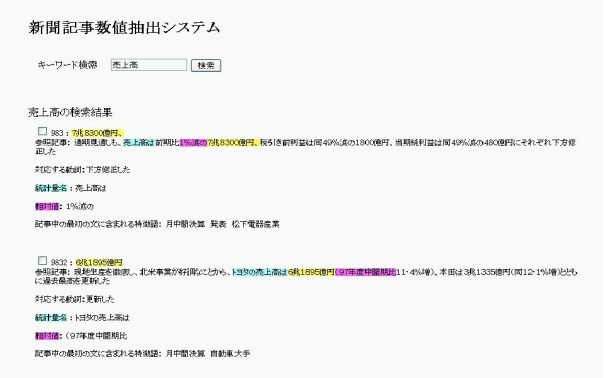

図 6 数値による情報検索システム

7. おわりに

本研究では, 数値情報に着目し, 動向情報における統計量名, 値, 相対値の自動抽出及び評価を行った．統計量名と統計量の相対値の抽出結果は，一部のトピックのみしか良い再現率は得られなかった．統計量名の抽出に関しては，問題点が明確されたため，抽出手法の改良を行う必要性があり，統計量の相対値は，統計量値の相対値の有無の判定の実装，そして相対表現の推定の精度の向上が必要である．今後は，統計量名と統計量の相対値の再現率と適合率の向上，そして新聞記事に存在する数値情報の関係性の特定を行う．

参 考 文 献

1) 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， JSAI2005 (2005).
2) 難波 英嗣，国政 美伸，福島 志穂，相沢 輝昭，奥村 学 : 文書横断文間関係を考慮した動向情報の抽出と可視化, IPSJ-NL168 (2005).
3) 藤畑 勝之，志賀 正裕，森 辰則 ：係り受けの制約と優先規則に基づく数量表現抽出，自然言語処理研究会報告 2001-NL-145, 情報処理学会, (2001).
4) 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に関するワークショップの提案，情報処理学会自然言語処理研究会，2004-NL-164(15)，pp.89-94, (2004)
5) 動向情報の要約と可視化に関するワークショップ, http://must.c.u-tokyo.ac.jp/
6) 毎日新聞社，CD-毎日新聞 98 年版
7) 毎日新聞社，CD-毎日新聞 99 年版
8) 係り受け解析器「Cabocha」, http://www.chasen.org/taku/software/cabocha/