2H5-9

新聞記事の数値による情報検索システムの提案と実装

Trend Information Extraction with Numeral Information from NewsPaper Articles

杉浦　隆博[*1]　吉田　稔[*2]　山田　剛一[*1]　増田　英孝[*1]　中川　裕志[*2]
Takahiro Sugiura　Minoru Yoshida　Kouichi Yamada　Hidetaka Masuda　Hiroshi Nakagawa

[*1]東京電機大学大学院工学研究科
Graduate School of Engineering, Tokyo Denki University Graduate School

[*2]東京大学情報基盤センター
Information Technology Center University of Tokyo

We propose a new IR system that provides "search by numbers". The system extracts numerical values (e.g, "￥5,000,000") from news articles, along with their attribute names (e.g., "business profits") and relative expressions (e.g., "50%" increase). The extraction algorithm uses the results of dependency parsing, hand-written rules for suffixes (e.g., "Yen" and "%" indicate numerical values) , etc. The extracted values are used to rank the articles from a particular point of view,e.g., the amount of company's profit.

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．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]．そして、動向情報のほとんどは数量表現によって表すことが可能である．

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シェア」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シェアである、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，本稿では数値情報の関係性を抽出する前段階として，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．

2. 動向情報コーパス

本システムでは「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 (通称：MuST コーパス）を参考にし、数値情報と関連する情報の抽出を行っている．

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
　<pro ref="前年比" id="9801220"> 同 </pro>
　<rel type="prop">3.1 ポイント </rel> 低い
　<val>82.7%</val>
　となった
</unit>.

タグの詳細な仕様に関しては表 1 に記す.

本稿では，数量表現に関係する情報として，統計量の名前，値，値の相対表現の自動抽出を行う.

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.
<name>	統計量の名前を示す.
<par>	出来事をおこした主体，出来事の一部となる事物など統計量名のパラメータを示す.
<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試みる．ここでいう数値

連絡先: 杉浦隆博，東京電機大学大学院工学研究科情報メディア学専攻，東京都千代田区神田錦町 2 丁目 2 番地，03-5280-3281 ext 2843，03-5280-3592，sugiura@cdl.im.dendai.ac.jp

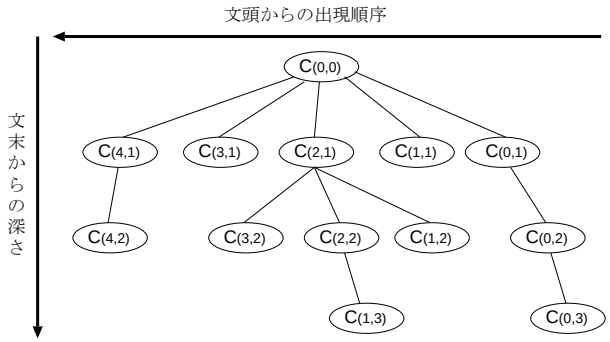

図 1: 依存構造木

情報とは, 数値と数値に関連する統計量名と相対値のことである．本手法では，まず Cabocha[8] による係り受け解析を行い，その結果を依存構造木の形に変換する．依存構造木の各ノードを $C_{(w,d)}(w =$ 文頭からの出現順序　$d =$ 文末からの深さ) とし，文末を $C_{(0,0)}$ とする．このときの依存構造木は図 1 となる．深さ d が大きいほどそのノードの出現順序が早いことを示す．また，複数のノードが同じ深さであるとき w が小さいほど出現順序が早くなる．同一の深さでないとき，ノード間の w の比較は意味を持たない．

本研究では、この依存構造木を用いて数値と関連する統計量名と相対値の特定を行うことになる．

3.1 数値の抽出

新聞記事を構文解析後に，数値とその数値の単位の抽出を行う．ここでいう数値とは統計量の値の候補となるものである．数値の特定には，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，数値ノードとする．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．

3.2 統計量名の特定

本研究では，依存構造木中にある数値ノードを手がかりにして，統計量名の特定を行う。数値ノードと繋がりを持つノードを辿り，統計量名と判定できた時点で抽出を終了する．

図 2 が全体の処理の流れである．文中にある全ての数値ノードを対象に，統計量名の探索を行う．まず，数値ノードの子要素が統計量名であるか判定し，統計量名ではない場合数値ノードの親ノード方向への統計量名の探索を行う．

親ノード方向への統計量名の探索の流れは図 3 に示す．まず，親ノードが統計量名であるか判定を行い，統計量名でなければ動詞または文末となるまで親ノードを辿る．動詞または文末となる親ノードが見つかったとき，そのノードの子ノード方向への統計量名の探索を行う．

子ノード方向への統計量名の探索の流れは図 4 に示している．ノード $C_{(i,j)}$ を元に，$C_{(i,j)}$ と同じ深さで一つ前に出現するノード ($C_{(i-1,j)}$) から，一番右の子ノードとなる $C_{(v,j)}$ までの統計量名の判定を行う．

どの数値ノードに対しても，統計量名を特定できた段階で統計量名の探索を終了する。探索を最後まで行った時点で，統計量名を特定できていなければ，その数値ノードに関しては「統計量名なし」となる．

統計量名の判定は図 5 に示す．統計量名の判定では，数値の単位と対応する統計量名が対象ノードに含まれるか判定し，対象ノードが統計量名を含んでいる場合そのノードを統計量名とする．単位と統計量名の対応情報は，毎日新聞の 98 年版と 99 年版から人手で抽出した情報であり，その一覧が表 2 となっている．

3.3 統計量の相対値の特定

基本的な処理の流れは統計量名の特定と同様である。相対値の特定に関しては, 文節中に「%」,「割」,「前期比」などの比率を表す語を含むものに着目し、これを統計量名の相対値とする．

S = {$C_{(w,d)}$ ｜ 名詞-数である文節 }
foreach $C_{(w,d)} \in S$ {
　//$C_{(v,d+1)}$ は $C_{(w,d)}$ の一番左の子ノード
　//$C_{(v',d+1)}$ は $C_{(w,d)}$ の一番右の子ノード
　for($n = v; n >= v'; n--$){
　　if (統計量名の判定 ($C_{(n,d+1)}$)) then {
　　　return $C_{(n,d+1)}$;
　　}
　}else{
　　//$C_{(v'',d-1)}$ は $C_{(w,d)}$ の親ノード
　　return 親ノード方向への統計量名の探索 ($C_{(v'',d-1)}$);
　}
}

図 2: 数値に関係する統計量名の探索

Node 親ノード方向への統計量名の探索 ($C_{(x,y)}$){
　if(統計量名の判定 ($C_{(x,y)}$)){
　　return $C_{(x,y)}$;
　} else {
　　for($n = y; n >= 0; n--$)){
　　　//$C_{(v(n),n)}$ は，$C_{(x,y)}$ から $C_{(0,0)}$ へ
　　　//辿った時の各ノード
　　　if($C_{(v(n),n)}$ が動詞または文末であるか）{
　　　　return 子ノード方向への
　　　　　統計量名の探索 ($C_{(v(n+1),n+1)}$);
　　　}
　　}
　}
}

図 3: 親ノード方向への統計量名の探索

Node 子ノード方向への統計量名の探索 ($C_{(i,j)}$){
　//$C_{(v,j)}$ は $C_{(i,j)}$ と同じ深さの一番右の子ノード
　for($m = i - 1; m >= v; m--$){
　　if(統計量名の判定 ($C_{(m,j)}$)){
　　　return $C_{(m,j)}$;
　　}
　}
　return null;
}

図 4: 子ノード方向への統計量名の探索

```
boolean 統計量名の判定 (C_(w,d)){
        if(C_(w,d) が表 2 の対応表現を含むか){
            return true;
        } else {
            return false;
        }
}
```

図 5: 統計量名の判定

表 2: 単位と統計量名の対応表

単位	対応する表現
円	売上高，売り上げ, 販売額, 販売高，利益，損失, 消費支出, 赤字, 連結決算, 費用，累損，負債市場, 歳入
台	出荷台数, 販売台数, 生産台数，　自動車生産, 国内生産, 海外生産, 海外販売, 出荷実績, 累計, 加入電話, 加入数, 規約当事者数, 増加数
%	シェア, 市場, 国内物価指数, 状況判断 DI, 一致指数, 普及率, 平均消費支出
単位なし	状況判断 DI, 物価指数, 卸売物価, 輸出入物価
ケース	出荷数量 出荷量 総量 発泡酒 ビール
度	気温

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 MuST コーパスを正解データとし，MuST コーパスに含まれる 1998 年から 1999 年までの毎日新聞の 581 記事を評価対象とする．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△，抽出できなかったものを×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 27 トピック毎に評価している．

表 3 が抽出結果全体の再現率の評価結果である．表 4 は特に評価が良かった統計量名と相対値の上位 3 トピックの再現率であり，表 5 は評価結果下位 3 トピックの再現率である．

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	35.1%	28.9%
再現率 (△を含む)	42.3%	29.2%

表 4: 上位 3 トピック

順位	統計量名	再現率	統計量の相対値	再現率
1	ソニー	95.0%	ソニー	84.4%
2	商業販売統計	76.5%	百貨店の売上高	83.3%
3	エアコン	76.4%	商業販売統計	82.3%

表 5: 下位 3 トピック

順位	統計量名	再現率	統計量の相対値	再現率
1	ガソリン	9.6%	景気予測	3.1%
2	長野五輪	10.0%	総合電機 3 社	11.3%
3	物価	16.3%	住宅	12.0%

4.2 適合率の評価

適合率の評価対象も、再現率と同様の 1998 年から 1999 年の毎日新聞の 581 記事である。適合率では、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．

表 6 が抽出結果全体の適合率の評価結果である．表 7 は特に評価が良かった統計量名と相対値の上位 3 トピックの適合率であり，表 8 は評価結果下位 3 トピックの適合率である．

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「人口」,「長野五輪」,「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」とい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全般的に低い適合率となっている．

表 6: 適合率の評価

	統計量名	統計量の相対値
適合率	57.1%	37.8%

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」,「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．

表 7: 上位 3 トピック

順位	統計量名	適合率	統計量の相対値	適合率
1	ソニー	94.1%	ソニー	70.6%
2	商業販売統計	82.5%	商業販売統計	66.7%
3	百貨店の売上高	80.7%	百貨店の売上高	64.9%

表 8: 下位 3 トピック（統計量名のみ）

順位	統計量名	適合率
1	長野五輪	24.7%
2	人口	31.7%
3	為替レート	33.3%

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
……				
	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．

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、我々が数値情報と統計量名の出現傾向の調査に利用するものであり，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．数値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．

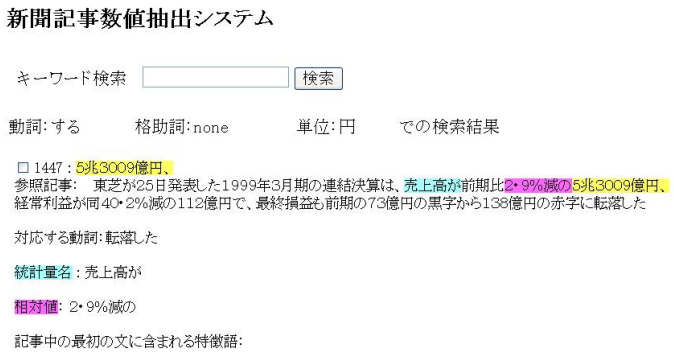

図 6: 数値による情報検索システム

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，問題点が明確化されたため，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，JSAI2005-1E3-03 (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，情報処理学会　自然言語処理研究会研究報告 2005-NL-168,pp67-74 (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，情報処理学会　自然言語処理研究会研究報告 2001-NL-145,pp119-125, (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/