新聞記事中の統計量名の学習による自動抽出

村田 一郎 †　　**森 辰則** ‡

† 横浜国立大学 大学院 環境情報学府　　‡ 横浜国立大学 大学院 環境情報研究院

E-mail: {ichiro,mori}@forest.eis.ynu.ac.jp

1　はじめに

製品の売上高や景気の情勢などの動向情報を様々な形で要約，可視化しようとする試みが広がっている [3]. これらの研究は，人間が処理するには困難な膨大な情報を扱いやすくし，統計データとしてまとめられていない情報へのアクセスが可能になるという点で有用である．

動向情報の要約と可視化を行なうにあたってまず必要なことは，文書から統計量を自動的に抽出することである．統計量は図 1のように，‘統計の調査方法’ と ‘値’ の組からなると考えられる．‘統計の調査方法’ は文書中には陽に現れず，それを指し示す表現が文書中に出現する．例えば，

> 「国内自動車メーカー大手５社は２１日，
> １９９７年の生産実績を発表した．」

という文において，‘統計の調査方法’ は「国内自動車メーカーの１９９７年の自動車の生産台数」である．しかし文中に現れている表現は「国内自動車メーカーの１９９７年の生産」であり，「自動車の」と「台数」という表現は現れていない．このように文書中での ‘統計の調査方法’ の表現は１通りとは限らず，様々な現れ方をする．‘統計の調査方法’ を指すこれらの文書中の表現のことを統計量名と呼び，本研究ではこの統計量名を抽出することを目標とする．なお，動向情報の要約と可視化に関するワークショップ (MuST) [1] ではすでに統計量に関する文にタグづけがされたコーパスが利用可能であるが，本研究ではそのコーパスですでに与えられているような情報を自動的に抽出することを目標とする．したがって，処理の対象はタグなどの情報が付加されていない文章になる．

統計量名を自動的に抽出するには，語の並びや前後関係などの規則性を見い出して抽出規則を構築し，その規則に対象文書を当てはめる方法が基本的であると思われる．しかし統計量名は文書中に様々な形で現れるため，それらの抽出規則を人手で網羅することは難しい．そこで，機械学習手法を用いて抽出規則の構築を自動化することを考える．そのためには，文書中に現れる統計量名の情報を学習データとして与えなければならないが，統計量名がどのような構造をしているか，つまり ‘統計の調査方法’ がどのように記述されるかを考えなければうまく学習データを作ることができない．本稿ではこの ‘統計の調査方法’ の構造の捉え方と，その問題点について考察を行なった．

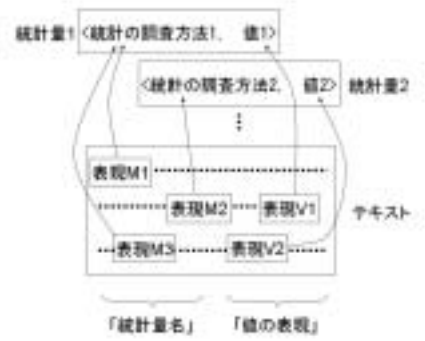

図 1: 統計量の構造

2　先行研究

統計情報の抽出に関して，斉藤ら [2] は数値の周りの言語パターンを調べ，それを当てはめることで統計量の抽出を試みている．また，藤畑ら [4] は数値に対する係り受けの制約を考察し，それに基づく優先規則を用いての情報抽出を提案している．いずれの研究でも統計量名は数値と関連のある名詞であるとされているが，どこまでを統計量名として抽出すれば十分かということは考慮されていない．本研究では統計量名を構成する表現は何かを検討し，統計量名として抽出すべき部分を明らかにする．

3　統計量名の自動抽出

3.1　統計量名の定義

統計量名とは，ある統計量の値がどのように統計を取って得られたものかを文章中で表現している部分である．単純なものを考えれば，「出荷量」「生産量」「加入台数」などが統計量名になる．しかし，それだけでは何の数量か，いつの時点での数量か，などということが分からず，統計の調査方法を特定するための表現としては不十分である．そこで統計量名としては，統計の調査方法

を説明する語句を過不足なく取り出す必要がある．例えば，

> 「携帯電話の加入台数は昨年１年間では１０５７万台増加し，NTT ドコモグループ９社だけで７０４万５０００台，６７％を占めていることが郵政省のまとめで分かった．」

という文において「１０５７万台」という値に対しては，「携帯電話の加入台数の昨年１年間の増加数」というものが統計量名として過不足のない表現となる．これを「加入台数の増加数」にしてしまうと不足であるし，また「NTT ドコモグループ９社」や「郵政省のまとめで」はこの統計量の値とは関係のない表現なのでつけるべきではない．以上のことを踏まえて，本稿では統計量名を「ある統計量の値がどのように統計を取って得られたものかを説明するのに必要かつ十分な文章中の表現をまとめたもの」と定義する．

文章によっては統計量名が複数の文に分かれて出現したり，必要な表現が文章中に出て来なかったりする．例えば，

> 「国内自動車メーカー大手５社は２１日，１９９７年の生産，販売，輸出，海外生産実績を発表した．４月以降の消費税率アップによる消費不振で３社の国内販売が落ち込んだが，円安と欧米の好景気を追い風にした輸出でカバーし，国内生産は５社とも前年を上回った．」

という文章では「１９９７年の国内自動車メーカー３社の自動車の国内販売台数」という統計量名が考えられるが，これを構成する表現は２つの文に分かれている．そのため統計量名の抽出は文を超えた広い範囲で考える必要がある．また，「自動車の」という表現は文章中に出て来ていない．文章中に出て来ない表現を扱うには文章の意味を高いレベルで理解することが必要であり，それには文章外の知識が必要と考えられるのでその部分は別の研究にゆずり，本稿では文章中に現れる表現のみを対象として考える．

3.2 統計量名の抽出方法

統計量名は文章中で単語の連続として出現するとは限らず，離れて出現する場合が多い．例えば，

> 「国内のビール大手５社は１３日，１月の課税出荷数量を発表した．全体の数量は３０５６万４０００ケースで，前年同月比１２・５％増と好調な滑り出し．特に首位のキリンビールが同２７％増の１２６９万１０００ケースと３２カ月ぶりの大幅な伸びを記録した．」

という文章では，「１月」「課税出荷数量」「全体の数量」「キリンビール」という表現が組み合わさって統計量名を構成している．そこで，統計量名はこのような１つ１つの表現から作られると考えて，それぞれを必要十分に抽出する方法が必要である．これら１つ１つを統計量名の要素と呼ぶことにする．

統計量名の要素を抽出するには，単語の表記や前後の情報などをもとに抽出規則を作り，それとのマッチングによって行うことも考えられるが，統計量名は文章中に様々な形で現れるため人手でそれらすべてを与えることは困難である．そこで，機械学習手法により抽出規則構築の部分を自動化する方法を考える．具体的には，文章を単語や文字などの単位に分け，その１つ１つについて前後の品詞情報などを用いてそれらが要素であるか否かを判定する２値分類問題とすることで，機械学習による抽出規則の構築が可能となる．

しかしこのように統計量名の要素を抽出すると，それぞれが独立に取り出されてしまうため要素間の関係が分からず，そのままでは統計量名としてまとめることができない．例えば，

> 「大手自動車メーカー５社が１８日発表した５月の国内生産の実績によると，日産自動車は前年同月比２２・８％減，トヨタ自動車は同２０・４％減となった．」

という文では，「５月」「国内生産」「日産自動車」「トヨタ自動車」が要素であるが，それぞれが結び付いて「５月の日産自動車の国内生産」「５月のトヨタ自動車の国内生産」という２つの統計量名ができる，と判断するのは要素の抽出とは別に考えなければならない．そこで，統計量名の抽出は以下の２つのタスクに分けて考えることができる．

- 文章中から統計量名の要素となるものをすべて取り出すタスク
- 取り出された要素をつなげて１つの統計量名を作るタスク

また，ここまでで取り出された統計量名は単なる要素の組合せであるが，これをもとに要約や統計情報の可視化を行なおうと考えた場合は，それぞれの統計量名の中でどれとどれが同じものか，という同一性判定を行なう必要があると考えられる．よって統計量名に関するタスクは，

- 統計量名を分類し，統計の取り方が同じものを判別するタスク

を加えた３つのタスクをもって完結すると考えられる．それぞれに課題が考えられるが，本稿では１つ目のタス

クを行なう上で必要な統計量名の要素の決定，及びそれと３つ目のタスクとの関係について考察する．

3.3 統計量名の要素

3.3.1 統計量名の内部構造

統計量名は統計の調査方法を特定するための様々な要素から成り立っているが，これらの要素は主要なものとそうでないものに分けて考えることができる．例えば，

> 「１９９８年のビールの出荷数量は，発泡酒を含めた総市場で前期に比べて０・７％減少した．」

という文において，統計量名は「１９９８年の発泡酒を含めた総市場でのビールの出荷数量」となるが，この中で主要なものは「出荷数量」であると考えられる．なぜなら，「出荷数量」はこの統計量をどのように数えたかを表現しており，統計の調査方法を説明する最も重要な語句だと言えるからである．そこでこれを１つの要素とし，後の「ビール」「１９９８年」などは主要素を説明する要素として考える．ただし，主要なものとして「ビールの出荷数量」を１つの要素と見ることもできるし，「発泡酒を含めた総市場でのビールの出荷数量」までを１つの要素とする考え方もあり得る．しかし後のタスクで統計量名どうしの同一性判定を行なうことを考えると，これは「出荷数量」という要素の説明として「ビール」「１９９８年」などがあると捉えた方が利用しやすい．そこで，主要な要素については細かく区切って考えることにする．その他の要素も，同一性判定の際に有用であるように分け方を考える．

統計量名には，「出荷数量」のように統計の調査方法に基づき構成的に作られたものと，「景気動向指数」のように統計の調査方法に対する名前づけではあるが，調査方法に基づいていないものがある．前者の方は，統計量名の内部構造を見るとどのような統計の調査方法を利用したかが分かる．この内部構造は次のような形をしていると考えられる．

> 条件＋対象＋数え方（＋比）

例えば，「１９９８年の発泡酒を含めた総市場でのビールの出荷数量」という統計量名において，「数え方」にあたるのは「出荷数量」であり，「対象」は「ビール」，「条件」は「１９９８年」と「発泡酒を含めた総市場」になる．また，割合などは何らかの具体的な数値を別の基準で表したものなので，具体的な数値を表す統計量名に続けて「比」の表現が現れると考えられる．例えば，「ビールの出荷量シェア」は「ビールの出荷量」を「シェア」で表すという形になっている．

この内部構造において，「数え方」（あるいは「比」）はその統計量名の主要な要素となり，他は説明のための要素となる．ただし「条件」よりは「対象」の方が統計の取り方として重要であると考えられる．

一方，後者の方は統計量名の内部構造を見てもどのような統計の調査方法を利用したかは分からない．このようなものは決まった内部構造は持たないと考えられる．

3.3.2 要素のタグセット

MuST のコーパスにおいては， name, date, dur, par, cond 要素が本稿における統計量名の要素となり得る．しかし前節で述べたような要素の分け方を考えると，これらのタグで囲まれた部分は本稿で考える要素とは必ずしも一致しない．例えば，

```
<name> ビールの出荷数量 </name>
<name> 発泡酒市場におけるシェア </name>
<name> 企業の景況感を示す業界判断指数 </name>
<name> ポケットベルの加入台数 </name>
```

のようなものは，統計量名の要素としてはもっと細かくするべきである．このため MuST コーパスをそのまま学習に使うことはできない．そこで学習データとして統計量名を与えるために，前節の議論を踏まえて統計量名として必要な細かさを考え，以下のようなタグセットを考案した．

head 統計量の値の最も基本的な数え方．「数」「量」「価格」など．

prop 統計量の値が割合などで表されている部分．「シェア」「倍率」など．

foot head がどのように生じたものかを表す部分．「出荷量」における「出荷」，「生産台数」における「生産」など．「出荷された数量」のように名詞の連続にはならない場合もある．

desc prop が何の割合かを説明する部分．「出荷量シェア」では「量」は head ではなくこれになる．

def 定義された式にしたがって計算された統計量の値．「景気動向指数」など．

obj 統計量の値が対象とする部分．「ビール」「自動車」など．

time 統計量の値を集計した期間を表す部分．１日でも期間とみなす．

range 統計量の値を集計した範囲を表す部分．「国内」「ビール市場」など．

add その他に統計量の値につけられる条件の部分．「合計」「平均」など．

この中で，head から def 要素までは統計量名の主要な要素となる． 前節で考えた構造に対応させると，head, foot は「数え方」，obj は「対象」，time, range, add は「条件」になる．prop は「比」であり，prop が現れた時だけ desc が「数え方」に当たる部分を表す．

ただしこのタグセットは，ビールの出荷数量や車の生産台数など統計量のある時点での値について主に述べられている文章を対象にして考えたもので，地震の発生や台風の上陸など，出来事に関して主に述べられた文章についてはうまく適用できないと考えられる．例えば，

「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した．」

という文では統計量名にあたる部分を選ぶことができない．(これは動向情報ではあるが，そもそも統計量が何であるかがはっきりとしない)

3.4 統計量名を要素に分ける際の考察

3.4.1 統計量名の主要な要素

「ビールの出荷数量」という表現においては，「出荷数量」が主要な要素になると考えられるが，「出荷」が「数量」を説明する表現であるととらえてそれぞれを分けることもできる．これまで「出荷数量」を１つの要素と考えてきたのは，これが統計量名の「数え方」にあたるとすることに問題はないと判断したからである．しかしこれらを１つにしておくと，例えば，

「国内で出荷されたビールの数量」

という統計量名において，「出荷された」がどのような要素になるかが記述できない．また，後に同一性判定を行なう際に「出荷」が「数量」を説明する表現であるということを与えた方が有効であることも考えられる．そこで，3.3.2節で考案したタグセットではこれらを分けて扱うこととした．

3.4.2 統計量の性質の違い

「出荷数量」「生産台数」などは純粋な数であるが，「シェア」「指数」などはそれとは性質の違うものである．このような要素が主要なものとなる統計量名は，具体的な数を別の基準で表している値の説明になっていることもあれば，何らかの計算によって算出された別の意味のある数値の説明であることもあり，これらは純粋な数を表したものとは別の内部構造を持つと考えられる．このうち，統計の調査方法に基づいて構成されている統計量名は，

条件＋対象＋数え方（＋比）

という構造で捉えることができると考えたが，これが常に正しいかどうかはまだ検討中である．また，統計の調査方法に基づいていない統計量名についてもまた別に検討する必要がある．

3.4.3 統計量名を限定する表現の粒度

統計量名の構造は，主要な要素とそれを説明するための要素からなると考えられる．主要な要素を説明するものは様々な種類があり，全てを細かく分類することはできない．例えば，

「トヨタ自動車の国内での生産台数は４８
０万１３８４台であった．」

という文において，「トヨタ自動車」というのは「生産」した会社名，「国内」というのは「台数」を数えた範囲であり，これらは別の要素になると考えることもできる．しかしそのように考えていくと，非常に多くの種類の要素が必要になってしまう．これらの要素は後の同一性判定において必要なだけの分類があればよい．この例では，「トヨタ自動車」と「国内」はどちらも統計量の値を集計した範囲であると考えて同じレベルで比較しても問題ない．しかしどれだけの分類があれば十分かということはまだ検討する必要がある．

4 まとめ

学習による統計量名の自動抽出を行なうにあたって，学習データとして統計量名を与える際に必要となる統計量名の内部構造について考察し，その問題点を指摘した．

今後は，多くの分野に共通する統計量名の内部構造を研究し，それを用いて実際に機械学習手法による統計量名の抽出，およびその考察を行なっていく予定である．

参考文献

[1] NTCIR-5 MuST. NTCIR-5 Pilot Workshop 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/, 2005.

[2] 斉藤公一, 迫田昭人, 中江富人, 岩井 禎広, 田村直良, 中川裕志. 数値情報をキーとした新聞記事からの情報抽出. 自然言語処理研究会報告 1998-NL-125, 情報処理学会, 1998.

[3] 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. 自然言語処理研究会報告 2004-NL-164, 情報処理学会, 2004.

[4] 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.