新聞記事からの統計量の抽出

藤岡 篤史 † **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: {fujioka,ichiro,mori}@forest.eis.ynu.ac.jp

1 はじめに

ある製品の価格や売上状況，内閣支持率などの動向情報に対する関心に，要約や可視化，またそれらを組み合わせたマルチメディアプレゼンテーションで答える研究が行われている [4].

各種文書に現れる動向情報を集約してその要約と可視化を行う場合には，文書から**統計量**に関する情報を抽出する必要がある．例えば，

> 「大手自動車メーカーが２４日に発表した
> １０月の国内生産実績によると，トヨタ自
> 動車は１４万台と前年実績を上回った．」

という文においては，表現「１０月の国内生産実績」，「トヨタ自動車」から推定される「トヨタ自動車の１０月の自動車の国内生産実績」という**統計の調査方法**と，それに対応する**値**を表現する「１４万台」の組が**統計量**の抽出結果となる．本稿では，前者の文書中における表出を**統計量名**と定義し，その自動抽出を検討する．特に，動向情報の集約を念頭に置き，統計量名を成す構成要素を分類された部品として抽出することを目標とする．例えば，先の例を集約して「月別のトヨタ自動車の自動車の国内生産実績」という動向情報を得るためには，統計をとった月を可変として，「トヨタ自動車の自動車の国内生産実績」に関する統計量名と対応する値を抽出することが必要である．

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)[4] では，統計量に関する注釈付けがなされているコーパスが提供されており，様々な研究がそれを基盤としてなされている．本研究では，そのコーパスで与えられているものと同等な情報を自動的に抽出することを目的としている．また，統計量を構成するもののうち，値に対応する表現の抽出は，比較的容易にできると考えて，本稿では考察の対象からはずしている．

2 先行研究

統計情報の抽出に関して，斉藤ら [3] 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 [5] 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．村田ら [6] は記事に出現する表現の頻度などの情報をもとに動向情報の抽出を行なっている．しかし，一記事から一つの動向情報しか抽出しておらず，一記事中に出現する複数の動向情報に対して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が何であるかを検討し，その構成要素を種別毎に区別して抽出をすることを目標としている．

3 文章中の表現と統計量との間の関係

次の二つの例文を考えよう．

例文１ 「Ａビールが発表した３月のビール出荷量は，２００万ケースだった．」

例文２ 「４月のＡのビール出荷数量は，２２０万ケース．」

統計量については，どのような統計であるかを表す表現 (例えば，「４月のＡのビール出荷数量」) と対応する値を表す表現 (例えば，「２２０万ケース」) の組で現れている．本稿では特に複雑な構造を持つ前者に注目をする．さて，二つ例のいずれにおいても，「(月別の) Ａビール社のビールの出荷量」に言及している点で共通しているが，それぞれ，「３月」と「４月」の統計であると言う点が差異となっている．複雑な統計量を収集して動向情報として集約するためには，このような共通部分と差異の部分を区別できる必要がある．更に，同じ統計量でも，どこが共通部分となり差異部分になるかは，どのような軸で統計量を収集するかによって変わるために，その部分構造を適切な種類に区別して認識することが要求される．

一方で，統計量に関する情報が文章中に現れる際の表記の多様性についても考慮する必要がある．上記の例では，「Ａビール」と「Ａ」，「出荷量」と「出荷数量」のそれぞれが，同一の指示物を指し示しているが表記は異なる．

上記の各点に対応するために，我々は，二つの概念，**統計の調査方法**ならびに**統計量名**を以下のように定義・導入し，統計量の整理を試みる．

統計の調査方法 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なのかを示す概念．文章中に直接現れるものではない．(「３月のＡビール社のビール出荷量」に対応する概念)

統計量名 統計の調査方法を指し示すために文章中に表出する表現を分類して組み合わせたもの．例えば，後述の分類に従うと例文１の統計量名は <agent:「Ａビール」, time:「３月」, obj:「ビール」, foot:「出荷」, head:「量」 > となる．

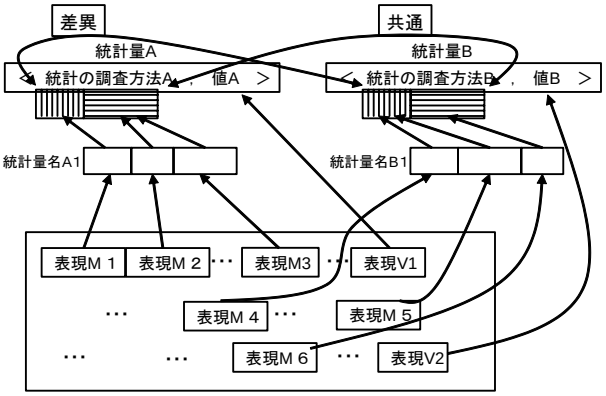

図 1: 文章中に現れる統計量の構造

統計量 ある「統計の調査方法」と，それに対応する値の組.

文章中に表出するときは，統計の調査方法を指し示す統計量名と，値を指し示す表現の組となって現れると考えられる．表現の多様性は，同一の「統計の調査方法」を指し示す「統計量名」の多様性に帰着して考える．図 1に上記の関係の概略を示す．なお，この図の示すとおり，統計量名の構成要素は文章中に分散していることもありえる点に注意されたい.

4 統計量と動向情報

4.1 統計量と出来事

加藤ら [4] は，動向情報はそれぞれの統計量に関する記述と，ある出来事に関する記述の２種類に分けられると述べている．以下に例文を示す.

例文３ 「乗用車の生産台数は４７３万５３７４台で，前年同期比１２・１%の大幅減少となった．」

例文４ 「台風７号は２２日午後１時ごろ紀伊半島に上陸し，大阪や京都などの近畿地方の主要都市を通過した．」

例文３は，「乗用車」という対象に着目し，「乗用車の生産台数は４７３万５３７４台である」という統計量に関する記述である．すなわち，ある主体や対象物に着目し，ある時点での値に注目した動向情報が統計量である．例文４は，「台風７号が紀伊半島に上陸した」という出来事に関する記述である．その年に幾つの台風が上陸したか等という統計的な記述ではなく，個々の記述である動向情報が出来事である．２種類の記述は動向情報にまとめられるが，統計量と出来事は別物と考え，本稿では統計量のみを扱うこととする.

4.2 統計量名の種類

統計量名は，少なくとも，以下の例に示す３種類に分類できる.

例文５ 「１９９８年度のパソコンの国内出荷台数は７３５万台と前年度比１０%増で，前年実績を上回った．」

例文６ 「１７日に中東のドバイ原油価格は１バレル当たり９・９８ドルであった．」

例文７ 「１月の景気動向指数は６２・５%となり，景気判断となる分かれ目である５０%を越えた．」

例文５は，何らかの動作によって生じた物の統計量を扱うものである．一方で，例文６はある物の状態や性質が統計量となっているものである．前者には，動作に関与する動作主等が統計量名の重要な一部として現れるが，後者は物の状態であるので，属性名が現れる．例えば，例文５はあるメーカが出荷したパソコンについての統計量となっている．一方，例文６では原油のそのものの属性を現す表現である価格が統計量である．これは統計の値の対象となるものそのものが統計量となるものである．例文７では「景気動向指数」が統計量名の主要部を成すが，これは外部で定義された何らかの式等に従って計算される方式に対する名前である．本稿では，以上の例文に示される統計量名の種類を，それぞれ，**動作型**，**属性型**，**定義型**と呼ぶことにする.

5 統計量名の自動抽出

5.1 統計量名の抽出タスクの構造

統計量名を構成する部品は文章中に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３０５万４０００ケースで，前年同月比１２５%と好調な滑り出し．」

という文章では，「１月」，「課税出荷数量」，「全体の数量」が組み合わさって統計量名を構成している．そこで統計量名がこのような１つ１つの表現から構成されていると考え，それぞれの表現を分類して抽出する方法が必要である．本稿では，これら１つ１つの表現を**統計量名の要素**と呼ぶことにする．統計量名の要素を個別に抽出した後は，適切な要素を組み合わせ，一つの統計量名を構成しなければならない．例えば，

> 「１８日に発表した５月の国内生産の実績によると，日産自動車は前年比２２・８%減，トヨタ自動車は同２０・４%減となった．」

という文において，「５月」，「国内生産」，「日産自動車」，「トヨタ自動車」が統計量名の要素であり，それらが結び付いて「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.

そこで本稿では，統計量名の抽出を以下の２つのタスクに分けて考える．

- 文章中から統計量名の要素となるものすべてを取り出すタスク
- 取り出された要素を組み合わせて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み合わせであるが，これを元に要約や統計情報の可視化を行おうと考えた場合，対応する「統計の調査方法」が何であるのかを復元し，同種の統計量を集める必要がある．その基本となるものが，

- 統計の調査方法が同じものを判定するタスク

である．なお，統計の調査方法自身は直接表現には現れないものであるから，それぞれの統計量名の中で共通部分と差異の部分を認識するタスクで代替することになると考えられる．

本稿の以降の部分では，１つ目のタスクに注目する．特に，統計量名の各要素がどのような分類になるかを考察し，それらの自動抽出を試みる．残りの二つのタスクについては，今後の課題としたい．

5.2 統計量名の内部構造

ここでは 4.2節で分類した３種類の統計量名について，それぞれの内部構造を考察する．

5.2.1 動作型の統計量名の内部構造

例文５の「１９９８年度のパソコンの国内出荷台数」という統計量名は，「１９９８年度」，「パソコン」，「国内出荷台数」という統計量名の要素から構成されている．「パソコン」という要素は統計を取る「対象」である．「出荷台数」は言い換えると「出荷された台数」であり，「出荷する」という「動作」と，「台数」という「数え方」で表されている．そして，「１９９８年度」や「国内」はこの統計量を限定する「条件」となっている．この例が示す通り，動作型の内部構造は，以下のような構造をしていると考えられる．

条件＋対象＋動作＋数え方

5.2.2 属性型の統計量名の内部構造

例文６の「ドバイ原油価格」という統計量名は，「ドバイ」，「原油」，「価格」という統計量名の要素から構成されている．「原油」は例文５の「パソコン」と同様に統計を取る「対象」である．しかし，「価格」は対象の量ではなく，対象の持つ「属性」の一つである．また，「ドバイ」は「原油価格」を限定する「条件」となっている．この例が示す通り，属性型の内部構造は，以下のような構造をしていると考えられる．

条件＋対象＋属性

5.2.3 定義型の統計量名の内部構造

例文７に関しては，統計量名は「１月の景気動向指数」であり，「１月」，「景気動向指数」という統計量名の要素から構成されている．ここで，「景気動向指数」は，何らかの計算方法によって定義されている量の名前に過ぎず，動作型や属性型と違い，内部構造を持たない．一方で，「１月」はこの統計量を限定する「条件」となっている．この例が示す通り，定義型の内部構造は，以下のような構造をしていると考えられる．

条件＋定義

5.3 統計量名の各要素を注釈付けするためのタグセット

各種表現を分類するために以下のタグセットを用意した．

- 動作型に関するタグ
  - **obj** 対象となる部分．「ビール」など．
  - **foot** 対象が受けた動作の部分．「出荷」「生産」など．
  - **head** 統計量の数え方．「数」「量」など．
  - **prop** 統計量の数え方が割合で表されている部分．「シェア」など．
- 属性型に関するタグ
  - **obj** 対象となる部分．「原油価格」における「原油」など．
  - **attr** 対象の属性を表す部分．「原油価格」における「価格」など．
- 定義型に関するタグ
  - **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．
- 「条件」に関するタグ (上記，統計量の各型に共通)
  - **time** 統計量の値を集計した期間を表す部分．
  - **locat** 統計量の値を集計した地域．
  - **agent** 会社名や機関名など．
  - **age** 年齢．
  - **add** 統計量の値に付加的につけられる条件の部分．「合計」「平均」など．
  - **range** 上記以外の統計を集計した範囲．

以下にタグを付与した例文を示す．

<agent id=”01,02”> トヨタ自動車 </agent> の <time id=”01,02”> １９９８年 </time> の <locat id=”01,02”> 国内 </locat><foot id=”01”> 生産 </foot><head id=”01”> 台数 </head> はわずかに減少したが， <foot id=”02”> 販売 </foot><head id=”02”> 台数 </head> は増加した．

6　統計量名の要素の自動抽出

テキストに出現するある種類の表現の抽出は，解析単位を一つのまとまりと同定する問題としてとらえることができる．本稿では，文字を構成単位としたチャンキング問題として，統計量名の要素の抽出を捉えることを考える．そして，比較的標準的な手法で我々の定義した各要素がどれくらいの精度で抽出できるかを調べる．具体的には，中野ら [2] で用いられている固有表現抽出の手法と同等の方法により統計量名の要素を抽出する．今回用いた素性は文字自身，文字種，品詞，単語，文節内素性，複合名詞主辞素性であり，文節内素性と複合名詞主辞素性は以下の通りである．

文節内素性 文節内に固有表現が存在すれば，最も先頭に近い固有名詞の品詞細分類を，固有名詞がなければ文節の先頭の単語を素性として用いる．中野らが新たに導入した素性である．

複合名詞主辞素性 連続する名詞が存在する場合，連続する名詞の最後の名詞を素性とする．

7　実験および考察

7.1　実験データ

実験には MuST コーパスで用いられている毎日新聞１９９８年，１９９９年の４８５記事をテキスト集合とし，統計量の動向情報である２３トピック (ガソリン，ビール業界など) に対し，5.3節で用意したタグを付与した．文単位に１０等分し，訓練データ９，評価データ１の比率で各要素の抽出に関する交差検定を行い，それらの平均の適合率，再現率で評価を行なった．

チャンキングには YamCha [1] を使用し，エンコーディング法には IOB2 を利用し，チャンキングの解析方向は左向き解析で行い，文脈長は対象文字の前後２文字ずつ，計５文字とした．

7.2　各タグの注釈付けの精度と考察

表 1に，各タグの自動抽出における適合率と再現率を示す．動作型の統計量名の主要素であり動作に対応する foot と，数え方に対応する head ，属性型の統計量名の主要素であり属性に対応する attr ，定義型の統計量名の主要素であり定義に対応する def に関しては適合率に関しては８０%以上，再現率でもほぼ８０%の精度であり，３種類の統計量名の主要素をある程度の精度で自動抽出できたと考えられる．しかし，動作型と属性型の統計量名の一部である対象に対応する obj は少し低い精度となった．これは対象となるものが様々であり，十分に学習できなかったと考えられる．条件部分であるそれぞれのタグに関しての結果については，適合率はある程度の結果と考えられる．一方，再現率に関しては地域をを示す locat が低い結果であるが，これも obj と同様に，地域の表現の種類が多かったことと，別途学習した学習データの少なさのためであると考えられる．固有表現抽出器の出力と組みあわせることにより，さらなる精度向上が期待されるが，今後の課題としたい．

表 1: 各タグの自動抽出精度

	obj	foot	head	prop	attr	def
適合率	76.5	80.1	86.0	74.0	80.7	84.7
再現率	64.4	79.3	85.4	76.4	74.6	79.3
	time	locat	agent	age	add	range
適合率	73.3	73.0	74.9	83.3	72.5	76.2
再現率	69.8	59.0	68.8	83.3	72.9	67.1

8　まとめ

本稿では，統計量名の種類とそれらの内部構造を定義し，統計量名の自動抽出に必要な，統計量名の要素の分類を機械学習手法により行った．今後は，統計量名の要素を１つにまとめる手法についての考察などを行っていく予定である．

参考文献

[1] http://cl.aist-nara.ac.jp/taku/software/yamcha.

[2] 中野桂吾, 平井有三.　日本語固有表現抽出における文節情報の利用．情報処理学会論文誌, Vol. 45, No. 3, pp. 934–941, 2004.

[3] 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.

[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.

[5] 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.

[6] 村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 塚脇幸代, 井佐原均.　MuST データを利用した自動動向調査システムの開発．言語理解とコミュニケーション研究会報告 NLC2005-119, 電子情報通信学会, 2006.