3H9-4

動向情報編纂のためのテキストからの統計量の自動抽出

Automated extraction of statistical expressions from text for information compilation

森 辰則[*1] 藤岡 篤史[*2] 村田 一郎[*2]
Tatsunori MORI Atsushi FUJIOKA Ichiro MURATA

横浜国立大学 大学院 環境情報研究院[*1]/環境情報学府[*2]
Graduate School of Environment and Information Sciences, Yokohama National Univeristy

In order to summarize trend information in document and visualize it, we have to have a method to automatically extract statistical information from document. Therefore, in this paper, we investigate automated extraction of statistical in formation, especially name of statistical information. We firstly classify the t ypes of name of statistics and examine the internal structure of them. Then, we propose a method to extract parts of name of statistics using a standard chunkin g algorithm. The experimental result shows that the proposed method is effective.

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 [加藤 04]．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，例文 (1)

(1) 「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.」

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．

2. 先行研究

統計情報の抽出に関して，斉藤ら [斉藤 98] は数値の周りの言語パタンを調べ，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 [藤畑 01] 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であるかということは考慮されていない．

一方，動向情報を扱った研究の多くは，動向情報の要約と可視化に関するワークショップ (MuST)[加藤 04] において報告されている．村田ら [村田 06] は記事に出現する表現の頻度などの情報をもとに，一記事から一つの動向情報の抽出を行っている．斎藤ら [斎藤 07] は，同一記事に現れる複数の統計量表現を，接尾表現に注目して抽出する手法を検討している．

これら先行研究に対し，本稿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点が異なる．

連絡先: 森 辰則，横浜国立大学大学院環境情報研究院，〒 240-8501 横浜市保土ヶ谷区常盤台 79-7, mori@forest.eis.ynu.ac.jp

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．

(2) 「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.」

(3) 「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」の統計であるとい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のかを示す概念．文章中に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組み合わせたもの．例えば，後述の分類に従うと例文 (2) 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．

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．上記の関係の概略を図 1 に示す．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得る点に注意されたい．

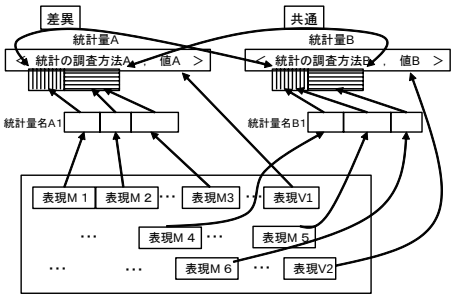

図 1: 文章中に現れる統計量表現の構造

4. 統計量名の種類

統計量名は，少なくとも，次例に示す 3 種類に分類できる．

(4) 「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.」

(5) 「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.」

(6) 「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 (4) は，ある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 (5) 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 (4) 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 (5) では原油のそのものの属性を表す価格が統計量である．例文 (6) 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．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．すなわち統計量名がこのような１つ１つの要素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．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．その基本は，

- 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，統計量名の各要素の分類を考察し，それらの自動抽出を試みる．残り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

ここでは節 4. 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．

動作型の統計量名の内部構造

例文 (4) 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」であるが,「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，動作型は，以下の内部構造を持つと考えられる．

条件＋対象＋動作＋数え方

属性型の統計量名の内部構造

例文 (5) 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 (4) 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，属性型は，以下の内部構造を持つと考えられる．

条件＋対象＋属性

定義型の統計量名の内部構造

例文 (6) 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，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，定義型は，以下の内部構造を持つと考えられる．

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

前節の考察に基づき，各種表現を分類し例文に注釈付けするために以下の XML 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.「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.

図 2 にタグを付与した例文を示す．なお，各タグは id 属性を持つ．これは，そのタグがどの統計量名に対応するものであるかということを管理する識別子であり，属性値中でカンマで区切られて示された各々の文字列がある統計量名に対応する．

```
<agent id="990000000_1, 990000000_2">トヨタ自動車</agent>の<time id=
"990000000_1, 990000000_2">１９９８年
</time>の<locat id="990000000_1,
990000000_2">国内</locat><foot id=
"990000000_1">生産</foot><head id=
"990000000_1">台数</head>はわずかに減少
したが，<foot id="990000000_2">販売
</foot><head id="990000000_2">台数
</head>は増加した.
```

図 2: タグ付与の例

6. 文字のチャンキングに基づく統計量名の要素の自動抽出

6.1 統計量名の要素の自動抽出

定義した各要素が，比較的標準的な抽出方法によってどれくらいの精度で抽出できるかを調べるために，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．チャンキングとは，任意の解析単位 (トークン) をある視点からまとめ上げていき，まとめ上げた固まり (チャンク) をそれらが果たす機能ごとに分類することであり，固有表現抽出などで用いられる．そこで，統計量名の要素の抽出には中野ら [中野 04] の固有表現抽出手法と同等の方法を用いる．6.2 節，6.3 節で同手法について述べる．

6.2 チャンクの表現方法

チャンキングを行う際，チャンクの状態をどのように表現するかが問題である．例えば各種先行研究においては，各トークンにチャンクの状態を示すチャンクタグを付与する方法が利用されている． 図 3 に「５月の国内生産」という文字列に対し，文字をトークンとして IOB2 法で符号化したチャンクタグの例を示す．第二カラム (「文字」) にある各文字のチャンク内の役割が，最終カラム (「タグ」) にチャンクタグとして示されている．チャンクタグは，対応するトークンのチャンク内での位置を表す記号と，チャンクの種類をハイフンで結んだものである．IOB2 では，チャンクの先頭トークンに B という記号を付与し，それ以降のトークンに記号 I を付与する．要素以外のトークンには O が付与される．

6.3 文字に対応する素性集合の分類に基づくチャンクの抽出方法

チャンキング問題は，各トークンに対しするチャンクタグの付与という分類問題に帰着される．Asahara ら [Asahara 03] の手法にならい，ここでは文字をトークンとする．文字列中の各文字にはその出現を特徴付ける素性を各種観点から割り当てることができる．例えば，図 3 において，第 1,2, 最終カラム以外の各カラムはそのような素性の例である．具体的には，素性として，文字自身，文字種，品詞，単語，文節内素性，複合名詞主辞素性を採用している．文節内素性とは，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ものである [中野 04]．また，複合名詞主辞素性とは，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ものである．

機械学習に基づく抽出手法を利用することを考えると，統計量名の要素の抽出規則の学習は，枠内の素性から対応するチャンクタグ (図 3 では B-locat) を得るような分類器を，学習事例と機械学習手法を用いて構成することに相当する．一方，未知の文における抽出の際には，各文字毎に枠内の素性集合を導出し，その素性集合を分類器に与えることによりチャンクタグを文末から文頭に向けて順次推定する．

位置	文字	文字種	単語	品詞	文節内素性	複合名詞主辞素性	タグ
	5	ZDIGIT	B·5月	B·名詞·副詞可能	5月	5月	B·time
i+2	月	OTHER	E·5月	E·名詞·副詞可能	5月	5月	I·time
i+1	の	HIRAG	S·の	S·助詞·連体化	※	※	O
i	国	OTHER	B·国内	B·名詞·一般	国内	台数	B·locat
i-1	内	OTHER	E·国内	E·名詞·一般	国内	台数	I·locat
i-2	生	OTHER	B·生産	B·名詞·サ変接続	国内	台数	B·foot
	産	OTHER	E·生産	E·名詞·サ変接続	国内	台数	I·foot

図 3: 素性集合に対する分類に基づくチャンクタグの推定

7. 実験および考察

7.1 実験データ

比較的標準的な手法であるチャンキング手法を用いることによって，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ために，統計量名の各要素の抽出実験を行った．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，統計量に関する動向情報である 23 トピックに対し，5.3 節で用意したタグを付与した文書を用いた．文単位とした 10 分割交差検定を行い，それらの平均の適合率，再現率で評価を行った．

なお，チャンキングには SVM に基づく汎用チャンカーである YamCha[Kudo 01] を使用し，チャンキングの解析方向は左向き解析で行い，各要素のタグの表現方法には IOB2 を利用し，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 各タグの抽出精度

各タグの自動抽出結果の評価を表 1 に示す．

表 1: 各タグの自動抽出に関する適合率と再現率

	obj	foot	head	prop	attr	def
頻度	978	672	417	275	500	168
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
頻度	2067	486	484	44	217	2362
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

動作型のタグについての考察

動作型の統計量名の主要素であり動作に対応する foot は適合率，再現率ともにほぼ 80%であり，数え方に対応する head に関しては適合率，再現率の両者が 85%以上であったため，動作型の主要素をある程度の精度で抽出できたと考えられる．動作型の統計量名に関しては，統計量名の要素がある程度一定の形で文書中に表出するため抽出精度が良くなったと考えられる．

属性型のタグについての考察

属性型の統計量の主要素である attr は，適合率はほぼ 80%ではあるが，再現率は 75%未満と低い結果である．属性には様々な表現があり学習が不十分であったためだと考えられる．そのため幅広い分野において属性の表現を学習させる必要がある．

また，対象に対応する obj に関しては適合率，再現率ともに低い評価結果となった．これは対象となるものが多く学習が不十分であったことが考えられる．

定義型のタグについての考察

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，再現率ともにほぼ 80%である．今回のコーパスでは「景気動向指数」,「国内総生産」,「平均消費性向」という表現ぐらいしか現れず，定義に対応する表現の数が少なかった．

条件のタグについての考察

条件に関するタグについては適合率がある程度の値となったが，再現率は全体的に低く，特に地域を示す locat が低い．これは attr や obj と同様に，文書に現れる地域の表現が多いのに対し，学習データの量が十分でなかったためであると考えられる．今後は，固有表現抽出器の出力と組み合わせることなどにより，精度を向上させる手法を検討すべきである．一方で，期間を示す time に関しては統計量に関連しない時間表現も抽出してしまっている．例えば，

> 「国内自動車メーカー大手 5 社は 1997 年の生産，販売，輸出実績を発表した．4 月の消費税率アップによる消費不振で国内販売が落ち込んだ.」

という文章において「4 月」という表現は統計量名の要素ではないが，このような時間表現も抽出をすることがある．

8. おわりに

本稿では，動向情報の要約と可視化を背景に，新聞記事からの統計量の抽出を目的とし，統計の調査方法と統計量名を定義することで，統計量名の抽出を検討した．動作型，属性型，定義型の 3 種類の統計量名の内部構造を定義し，それぞれの要素の抽出実験を行った．

その結果，統計量名の構造を分類することで，汎用チャンカーに基づく標準的な抽出方法を用いても，ある程度の精度で統計量名の要素を抽出できることがわかった．

今後の課題としては，統計量名の抽出で残された 2 つのタスク，すなわち，

- 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
- 統計の調査方法が同じものを判定するタスク

を実現する手法の検討がある．また，動向情報の要約や可視化を自動化するためには，統計量の値の抽出や，その値がどの統計の調査方法と組になるかを判定することも必要である．

参考文献

[Asahara 03] Asahara, M. and Matsumoto, Y.: Japanese Named Entity Extraction with Redundant Morphological Analysis, in *Proceedings of HLT-NAACL 2003* (2003)

[Kudo 01] Kudo, T. and Matsumoto, Y.: Chunking with support vector machines, in *Proceedings of NAACL 2001*, pp. 1–8 (2001)

[加藤 04] 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)

[斉藤 98] 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 98-NL-125, 情報処理学会 (1998)

[斎藤 07] 斎藤 悠, 河合 英紀, 土田 正明, 水口 弘紀, 久寿居 大: 新聞記事コーパスからの統計量表現自動抽出と共起関係ネットワーク構築, MuST「動向情報の要約と可視化に関するワークショップ」第二回成果進捗報告会論文集 (2007)

[中野 04] 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)

[藤畑 01] 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)

[村田 06] 村田 真樹, 一井 康二, 馬 青, 白土 保：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)