3K3-6

# 動向情報編纂のためのテキストからの統計量表現の自動抽出

## ― 統計量名の構成要素に関する組の同定 ―

## Automated extraction of statistic expressions from text for information compilation

## ― Identifying combinations of elements of statistic names ―

森 辰則[*1]
Tatsunori Mori

上野 史紀[*2]
Fuminori Ueno

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

In order to summarize trend information in documents and visualize it, we have to extract statistic information from documents. In our previous work, we reported a method to extract statistical names on an element-by-element basis. In this paper, we propose a method to identify combination of elements of statistic names automatically. First, it identifies the last element of each statistic name, and then it collects appropriate elements that appear up to the last element in order to combine them into a statistic name. The experimental result shows that the two-staged method is effective.

## 1. はじめに

ある製品の価格や売り上げ状況、内閣支持率などの動向情報に対する関心に対して、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、

(1)「大手自動車メーカーが24日に発表した10月の国内生産台数によると、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、「10月の国内生産台数」、「トヨタ自動車」という表現から推定される「トヨタ自動車の10月の自動車の国内生産台数」という統計の調査方法と、それに対応する値の表現である「14万台」の組が、統計量の抽出結果となる。藤岡ら[藤岡 07]は、前者の文書中での表出を統計量名と定義し、その自動抽出を検討している。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している。本稿では、藤岡らの研究を受け、文書中に散在している統計量名を成す要素を組み合わせて、一つの統計量名に同定するタスクを検討する。特に、第一段階として、統計量名を構成する最後の要素を同定し、次いで第二段階として、その統計量名を構成する最後の要素よりも前に出現する要素を集めて一つの統計量名にするという二段階手法を提案する。

## 2. 統計情報の抽出に関する先行研究

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、それを文章に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数値を対応する事物と組にしての情報抽出を提案している。しかし、いずれの研究でも統計量に対応する事物は数値と関連のある名詞であるとされており、それらをどこまで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察されていない。

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[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。

これらの研究に対して、藤岡ら[藤岡 07]は、統計量名を構成する表現が何であるかを検討し、その構成要素を種類ごとに区別して抽出することを目標とした研究を行っている。本稿は、この藤岡らのモデルを用いて抽出された統計量名の構成要素を組み合わせて、一つの統計量名として同定することを検討する。

## 3. 組同定の自動化

### 3.1 統計量名の要素の出現の仕方

図1のように、統計量は、基本的に統計量名の要素が幾つか連続して出現し、その直後に対応する値が出現する。一般には、一文章中に複数の統計量が出現し、要素によっては複数の統計量名で共有されることもある。

販売不振から家電メーカーの業績低迷の主犯格とされた<obj id="2_1, 2_2, 2_3, 2_4">エアコン</obj>の<time id="2_1">5月</time>の<locat id="2_1, 2_2, 2_3, 2_4">国内</locat><foot id="2_1, 2_2">出荷</foot><head id="2_1, 2_2">台数</head>が、昨年3月以来、1年2カ月ぶりに<time id="2_2">前年同月</time>比プラスに転じた。
（中略）
このため、<agent id="2_3, 2_4">東芝</agent>は<time id="2_3">6月</time>の<foot id="2_3, 2_4">生産</foot>計画を<time id="2_4">前年同月</time>比10%増に上方修正した。

上記文章より抽出される統計量名の要素の組

| id | obj | time | locat | foot | head | agent |
|---|---|---|---|---|---|---|
| 2_1 | エアコン | 5月 | 国内 | 出荷 | 台数 | |
| 2_2 | エアコン | 前年同月 | 国内 | 出荷 | 台数 | |
| 2_3 | エアコン | 6月 | 国内 | 生産 | | 東芝 |
| 2_4 | エアコン | 前年同月 | 国内 | 生産 | | 東芝 |

**図1 統計量名の要素の出現の仕方**

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

例えば、図 1 の文章には4つの統計量名が登場しているが、2行目に出現する表現「エアコン」は4つ全ての統計量名の要素になっている。一方で「5月」については1つだけの統計量名の要素になっている。つまり、統計量名を構成する要素の組を得るためには、各要素がどこまで後の文章に継承されているかを判断する必要がある。ここで、obj は対象、time は集計期間、locat は地域、foot は対象が受けた動作、head は統計量の数え方、agent は動作主である。詳細は[藤岡 07]を参照されたい。

図 1 の表のように、統計量名の各要素を種類ごとに分類して表現し、ある種類の要素について注目すると、その種類の要素について、新たに要素が出現しない限り、前の統計量名の同じ種類の要素が継承されることが多い。一方で、図 1[2-2]の head 要素のように後の統計量名に継承されない要素もある。

ある統計量名の側から考えたときには、その要素の種類毎に、上記のように過去に出現するどの要素を継承するかを決定する必要がある。これらのことを考慮して、統計量名の要素の組同定を実現する手法を次節にて提案する。

## 3.2 機械学習手法に基づき統計量名の要素を組として同定する手法の提案

### (1) 統計量名の要素を組として同定する手法

統計量名の各要素を組として同定するタスクは、例えば、図 1 から id 属性を取り除いた注釈付きのテキストを入力として受け付け、図 1 のように統計量名の要素の組を表す id 属性を付与することに対応する。プレインテキストに対して統計量名の各要素を注釈付けする手法については藤岡ら[藤岡 07]が検討しているので、本稿では、その手法の存在を前提としている。

具体的には、図 1 における表の一行分に相当するフレーム構造を準備し、obj や time といった各要素の種類に対応するスロットを、出現した適切な要素で埋めていく過程として捉えることができる。しかしながら、各統計量名について必ずしも可能な全ての要素の種類が文章中に出現するわけではなく、いくつの要素の種類から構成されているか不明である。そこで、まず基準点となる要素を決定し、次にその基準点の要素と組になる要素を全て取り出して、統計量名の組として同定する二段階の過程を考える。ここで基準点となる要素と組になる要素のことを「結びつく要素」と呼ぶ。

どの要素を基準点にするかについて考えると、ある特定の種類の要素が必ず出現するわけではないので、要素の種類を手がかりにすることはできない。そこで、各統計量名について、それを構成する要素のうち、文章中で最後に出現する要素を基準点とし、それ以前に出現する要素のどれと結びつくかを決定できればよいと考えた。

以上より、次の手続きに従って、統計量名の要素の組同定を試みる。ここで、一つの統計量名を構成する要素の組において、最後に出現する、基準点となる要素を「最終要素」と呼ぶ。

① 文章を先頭から走査し、出現する各要素に対して、「最終要素」であるか否かの判定を行う。
② 文章を再び先頭から走査し、以下の③から⑤までを新しい「最終要素」が出現しなくなるまで繰り返す。
③ 「最終要素」が出現した時、空のフレームを用意し、その「最終要素」を対応する種類のスロットに入れる。他の各スロットに対応する要素の各種類に対して、「最終要素」以前に出現した要素の中のどの要素と結びつくか、または、どの要素とも結びつかないかの判定を行う。結びつく要素があれば、その要素を抽出し、当該スロットに入れて保存する。
④ ③において保存された要素群を、1つの統計量名を構成している要素群として同定、出力をする。
⑤ フレーム上に保存されている要素を全て破棄し、走査を再開する。

### (2) 機械学習手法による実現

さて、上記手続きの中で①と③において、二段階の判定を行っている。①での判定を第一段階、③での判定を第二段階とする。本研究では、いずれの段階も Support Vector Machine (SVM) を用いた機械学習手法により、その判断を行う分類器を教師情報から自動的に獲得することを試みる。

第一段階の学習過程においては、図1のように id 属性が付与された学習用注釈付きコーパスを用いて、各要素が「最終要素」であるか否かを判定する分類学習を行い、分類器を生成する。運用過程である分類時には、要素の種類だけが注釈付けられ、id 属性による組情報が無い未知の文章に現れる各要素について、「最終要素」か否かの判定を行う。

また第二段階では、判断された最終要素の各々について、先に述べたフレーム構造を用意し、スロットに対応する各要素の種類毎に適切な要素を一つずつ決定する。要素のある一つの種類に注目すると、この過程は、i) テキスト中に現れる同じ種類の要素の中から、ある基準に従って候補群を集め、ii) その候補群の中から、現在の最終要素と適切に結びつく候補を一つ選択する、という手順からなる。候補群は基本的には最終要素よりも前に登場する、当該種類の要素の全てからなるが(後に述べる「手法 1」)、他の最終候補との位置関係によって候補を絞り込むことも考えられる(後に述べる「手法2」)。

複数の候補の中から適切な一つを選択する手法は機械学習によって実現できる。本稿では、図 2 に示す飯田ら[飯田 04]のトーナメントモデルを使用した。この手法は、複数存在する候補の中から最尤の候補を選択する過程を、複数の候補が参加するトーナメント戦(勝ち抜き戦)として捉えることにより、2つの候補間での勝負に還元するものである。学習される分類器は2つの候補を入力したときに、どちらがより優先される候補であるのかを判別する二値分類器である。未知の候補群からこの二値分類器を複数回適用することにより、トーナメント戦を行う。ところで、先に述べたように全ての種類の要素が文章中に必ず登場するわけではないので、その場合を想定する必要がある。本研究では、「どの候補とも結びつかない」ことを表す「なし」という候補も候補群に入れることによって、一回のトーナメントで解決することを試みる。

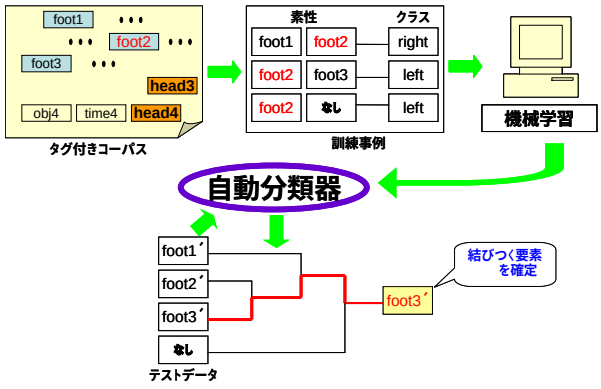


**図 2 トーナメントモデルによる結びつく要素の同定**

### 3.3 機械学習に用いた素性

前節で述べた2段階の処理それぞれについて、機械学習に用いた素性を以下に示す。

#### (1) 第一段階の分類器で用いた素性

以下の素性の組を事例としている。

[1-1] 当該要素の種類
[1-2] 一つ前に出現する要素の種類
[1-3] 一つ後に出現する要素の種類
[1-4] 当該要素の品詞情報
[1-5] 一つ前に出現する要素の品詞情報
[1-6] 一つ後に出現する要素の品詞情報
[1-7] 当該要素の表層表現
[1-8] 一つ前に出現する要素の表層表現
[1-9] 一つ後に出現する要素の表層表現
[1-10] 直後に統計量名の要素が出現することなく、数値情報が出現するならば1、そうでないならば0
[1-11] 次に登場する要素の種類が当該要素の種類と同じならば1、その他ならば0
[1-12] 直後に統計量名の要素が出現することなく、句点「。」が出現するならば1、そうでないならば0

[1-4]から[1-7]までの素性における品詞情報は、要素に現れた形態素ごとの品詞の頻度を素性としている。

素性[1-10]は、統計量名が述べられた直後に対応する値が登場しやすい、つまり、数量表現の直前にくる要素が統計量名の「最終要素」になりやすいという観察に基づいている。

素性[1-11]は、同一の種類の異なる要素がある一つの統計量名の要素にならないという性質を用いたものである。その場合、それぞれの要素は異なる統計量名の一部となるはずであるから、当該要素は最終要素になると考えられる。

素性[1-12]は、統計量名を構成する要素間に句点が入り込みにくいというに性質を用いている。つまり、句点の直前の要素は最終要素になりやすいという観察に基づいている。

#### (2) 第二段階での分類器で用いた素性

比較をする二つの要素の各々について、以下の素性の組を求め、それらを連結したものを事例としている。

[2-1] 当該要素の種類
[2-2] 当該要素の表層表現
[2-3] 当該要素の品詞情報
[2-4] 当該要素といま考慮している最終要素との間の距離
[2-5] 当該要素が、文章中で一番目の統計量名の要素である場合に1、そうでないならば0

素性[2-5]は、文章中で一番目に登場する統計量名の要素が、文章全体を通しての話題となりやすいという特性を表している。なお、どの候補とも結びつかない「なし」という候補に対しては、「素性がない」という素性を別に作って用いている。

## 4. 実験および考察

### 4.1 実験の設定

実験には MuST コーパス（2006 年版）を用いた[加藤 04]。MuST コーパスのうち、統計量の動向情報に関する 23 トピック、485 記事に対し、MuST で定義された注釈を取り除き、藤岡らが定義したタグが付与されている文書を用いた。

### 4.2 第一段階：最終要素同定に関する実験

4.1 節で準備した注釈付きコーパスに対し、統計量名の「最終要素」を同定する実験を行った。同コーパスのトピック毎に文書単位で 10 分割をし、交差検定により、適合率、再現率、F値により評価を行った。紙面の都合により、アルファベット順で先頭 2 つのトピックの各値と、トピックに亘る各値の平均値、標準偏差、最良値、最悪値のみを表 1 に示す。表1によれば、統計量名の「最終要素」の同定に対する精度はF値にして 0.9 程度となり、十分に高い精度を達成している。

**表1 第一段階: 最終要素同定の精度**

| トピック | 適合率 | 再現率 | F値 |
|---|---|---|---|
| AirConditioner | 0.952 | 0.924 | 0.938 |
| BeerIndustry | 0.838 | 0.895 | 0.866 |
| 平均値 | 0.902 | 0.899 | 0.899 |
| 標準偏差 | 0.056 | 0.072 | 0.057 |
| 最良値 | 0.974 | 0.995 | 0.984 |
| 最悪値 | 0.792 | 0.678 | 0.730 |

### 4.3 第二段階：組の同定に対する実験

組の同定の判定を行う対象である候補群は基本的には最終要素よりも前に登場する、当該種類の要素の全てからなる。これらを候補とした場合を「手法 1」としよう。ここで、一つ前の最終要素と、いま注目している最終要素との間に登場する要素について、「手法 1」ではあくまでも一つの候補としてのみ扱っており、優遇することはなかった。しかし、そのような要素は、いま注目している最終要素と結びつくことが多いので、それらについては、トーナメントモデルによる優先順位づけは行わず、「不戦勝」扱いとして優遇することが考えられる。この手法を「手法2」とする。

#### (1) 手法1

まず要素の種類それぞれに対して、2つの候補のうちどちらが適切な候補であるかを分類する分類器を構築したときに、どれくらいの精度で分類できるかを調べた。各表において、「1 対 1」と記されている。また、それらを複数回利用し、トーナメントを構成した後に、複数の候補群から一つの候補を選択すると、どれくらいの精度が実現できるのかを調べた。こちらは各表において、「トーナメント」と記されている。

いずれも、各トピックに対して文書単位に 10 分割交差検定を行い、精度の評価を行なった。結果を表2に示す。

**表2 第二段階: 組の同定の精度(手法1)**

| トピック | 1 対 1 | トーナメント |
|---|---|---|
| AirConditioner | 0.322 | 0.316 |
| BeerIndustry | 0.590 | 0.427 |
| 平均値 | 0.652 | 0.499 |
| 標準偏差 | 0.105 | 0.122 |
| 最良値 | 0.813 | 0.747 |
| 最悪値 | 0.322 | 0.198 |

表2に示す通り、トーナメントモデルを用いたときの、「最終要素」と結びつく要素の同定に対する精度は、0.2 から 0.7 程度となり、トピックによって精度に差が生じた。この理由については、以下の 2 つが考えられる。

1つ目として、トーナメントの段階において、どの候補とも結びつかないという意味の「なし」が正解である場合の多少が影響を与えていると考えられる。各トピックに対して、トーナメントの総数（①）に対し、その中で「なし」が正解である数（②）の割合（②/

①) を計算し、その割合とトーナメントモデルの精度について相関があるかをトピックに亘る相関係数を求めることにより調べた。その結果、相関係数は－0.436 となり、割合②/①とトーナメントモデルの精度の間に負の相関があることがわかる(有意水準 0.05)。このことより、「なし」が正解である場合が多いほど、正しい候補を選べていないことが分かる。

2つ目として、一つの統計量名の中に要素の種類が重複している場合による影響が考えられる。例えば次のような例がある。

(3)「1998年のビールの課税出荷数量をまとめ、発表した。(中略)。総市場ではキリンが2億8000万ケースとトップ。」

この場合の統計量名は、「1998年」、「ビール」、「課税」、「出荷」、「数量」、「総市場」、「キリン」の要素から構成されるのだが、これを要素の種類別に分けると、「課税」と「総市場」が{range}として分類されている。今回のトーナメントモデルでは、一つの統計量名に対して要素の種類が重複する場合を考慮していないので、精度が低下している原因と考えられる。

今回用いたコーパス中に、種類が重複した要素数は 3550 個あり、その 77%が range に分類されていた。このことより range に分類されている要素は、トーナメントモデルとは異なる別の処理により要素の組同定を行うことにより解決できると考えられる。

#### (2) 手法2

結果を表 3 に示す。表 2 と表 3 を比較すると、手法 2 の「優遇措置」を採用することにより、トーナメントの精度の平均値が 0.5 程度から 0.6 程度まで向上することが確認された。

**表 3 第二段階: 組の同定の精度 (手法2)**

| トピック | 1 対 1 | トーナメント |
|---|---|---|
| AirConditioner | 0.860 | 0.735 |
| BeerIndustry | 0.678 | 0.524 |
| 平均値 | 0.714 | 0.611 |
| 標準偏差 | 0.080 | 0.098 |
| 最良値 | 0.860 | 0.780 |
| 最悪値 | 0.550 | 0.420 |

### 4.4 注釈付けの揺れを修正したときの第一段階、第二段階に対する実験

前節までの実験において、コーパスにおける注釈付けの不統一が見受けられたので、これを統一した場合にどのように精度が変わるかを調べた。具体的には、AirConditioner, BeerIndustry のトピックに対して、コーパスに以下の2点の修正を施した。1つ目は、ある要素がいま注目している統計量名を構成する要素であるか否かの判断について、そう解釈してもしなくても文脈上はいずれも不都合がない事例があるが、それらについて「統計量名を構成する要素である」と判断をしなおした。2つ目は、種類が重複している要素に対して、種類毎に一つの要素にできるものをそのように修正した。

このコーパスを用いて第一段階、第二段階(手法 2)の実験を再度行った。結果を表 4 と表 5 に示す。表 1 と表 4 を比較すると、第一段階についてはあまり変化がなく、高い精度が保たれていることがわかる。一方で、表 3 と表 5 を比較した時に、二つのトピックいずれについてもトーナメントモデルによる組同定精度が上昇しており、第二段階目についてはコーパスにおける注釈付けの揺れを極力抑えることが有効であることが分かった。

### 4.5 二段階の過程を続けて実行した際の実験

第一段階と第二段階の過程を続けて実施した場合、各要素について、候補群の中から適切なものを選んでいるか否かについて適合率と再現率により評価を行った。結果を表6に示す。適合率、再現率共に 0.7 程度と二段階の各々の処理精度の積程度となった。

**表 4 修正後の第一段階の精度**

| トピック | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| AirConditioner | 0.944 | 0.934 | 0.939 |
| BeerIndustry | 0.871 | 0.856 | 0.863 |

**表 5 修正後の第二段階(手法 2)の精度**

| トピック | 1 対 1 | トーナメント |
|---|---|---|
| AirConditioner | 0.885 | 0.789 |
| BeerIndustry | 0.819 | 0.803 |

**表 6 二段階の過程を続けての実験**

| トピック | 適合率 | 再現率 |
|---|---|---|
| AirConditioner | 0.695 | 0.720 |
| BeerIndustry | 0.674 | 0.697 |

## 5. まとめ

本研究では、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名の抽出を目的とし、藤岡ら[藤岡 07]が定義したタグセットによって注釈付けされた要素を組み合わせて、一つの統計量名を作る手法を検討した。具体的には、第一段階として、統計量名を構成する最後の要素を同定し、次いで第二段階として、その統計量名を構成する最後の要素よりも前に出現する要素を集めて一つの統計量名にするという二段階手法を提案した。評価実験によれば、第一段階はある程度の精度で遂行できることがわかった。第二段階ではトピックごとに精度の差が生じたが、その原因を考察し、有効であると思われる改善方法を示した。

今後の課題としては、精度向上のために新たな素性の検討やコーパスの改訂が考えられる。また、統計量名の抽出の最後のタスクである「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。

## 参考文献


[飯田 04] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌, Vol45, No.3 (2004)

[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. 情報処理学会研究会報告, 2004-NL-164, pp.89-94 (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. 情報処理学会研究会報告, 1998-NL-125 (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 第二回成果進捗報告会論文集 (2007)

[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. 情報処理学会研究会報告, 2001-NL-164 (2001)

[藤岡 07] 藤岡篤史, 村田一郎, 森辰則. 新聞記事からの統計量の抽出. 言語処理学会第 13 回年次大会発表論文集, pp.119-122 (2007)

[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. 電子情報通信学会研究会報告, NLC2005-119 (2006)