統計量名の構造に関する一考察とその自動抽出

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 [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,

> 「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)[加藤 04] 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. 本稿では, 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 [斉藤 98] 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 [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に関しては, 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. また,MuST コーパスを用いることによって動向情報を可視化する研究 [山本 06] なども行われている. 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.

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).

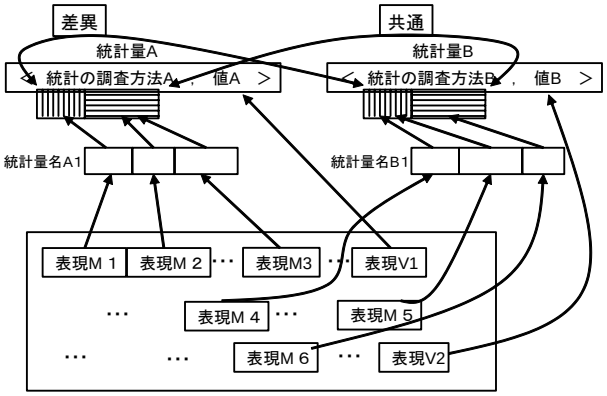

図 1: 文章中に現れる統計量の構造

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.

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 1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 [加藤 04] 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.

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. 詳しくは本稿末尾の付録を参照されたい.

```
<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. そこで本稿では, 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,6.4節で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タグを付与する方法が利用されている. チャンクの状態を表すタグ集合としては IOB1 法,IOE1 法,IOE2 法などが提案されているが,SVM を用いた固有表現抽出 [山田 02] においては, IOB2[Sang 00] と呼ばれるチャンクタグ集合を用いた場合が最も精度が良いと報告されている. 図 3に「日本の立場」という文に対して固有表現タグを付与した例を示す. ここで固有表現タグとは, チャンクタグと固有表現の種類をハイフンで結んだものである. IOB2 では, 固有表現の先頭トークンに B タグを付与し, それ以降のトークンに I タグを付与する. 要素以外のトークンには O タグが付与される.

文字	固有表現タグ
日	B - LOCATION
本	I - LOCATION
の	O
立	O
場	O

図 3: チャンクの状態を示すタグ

6.3 文字単位の素性展開

先行研究の多くはまとめ上げの単位として形態素を用いているが, この手法では形態素の境界が固有表現と一致しないと抽出できない. 例えば「神奈川県内で」を ChaSen を用いて形態素解析を行うと「神奈川 / 県内 / で」と分割されてしまうため「神奈川県」を地名として抽出できない. Asahara ら [Asahara 03] は, 文字をまとめ上げの単位として用いることによって, 形態素解析による単語の境界と固有表現の境界の不一致の問題を解消できることを示している. また, 文字そのものを素性として使用するため単語そのものを用いるよりも粒度の細かい情報を用いることができる.

形態素を単位とする場合と異なり, 文字には直接品詞情報を付加することはできない. そのため中野ら [中野 04] は各文字が属する単語と品詞に, その文字の単語中の位置に応じて Start-End 法 (以下 SE 法)[内元 00] に基づくチャンクタグを付与したものを素性として用いている. SE 法では形態素の先頭文字に対し B タグ, 末尾に E タグ, 内部に I タグ, 一文字からなる形態素に対しては S タグが付与される.

6.4 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

図 4は 6.2節,6.3節の手法により文字列を素性展開し, 対応するチャンクタグを付与した例である. ここで, 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 4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 4: 素性集合に対する分類に基づくチャンクタグの推定

ここで統計量名の要素を抽出するための手法を以下に示す. この手法は中野ら [中野 04] が固有表現を抽出するために用いている手法に, 複合名詞主辞素性を加えた方法である. まず, 学習の手順は以下の通りである.

step1 入力文に対し, 形態素解析及び文節区切りを適用する.

step2 各文字が属する単語や品詞情報に加え文節素性を展開する. 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. 中野ら [中野 04] が新たに導入した素性である. 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.

step3 各文字に対応する素性を入力とした時にそのチャンクタグを出力する分類器を機械学習により求める. 図 4のように前後各 2 文字分んを文脈とする時には,i 番目の文字に実線で囲まれた枠内の素性を対応させる. すなわち,i-2 番目から i+2 番目の文字自身, 文字種, 品詞, 単語, 文節内素性, 複合名詞主辞素性と i-2 番目と i-1 番目のチャンクタグを用いる. これらの素性の束と, 対応する文字に振るべきチャンクタグを組にしたものをタグ付きコーパスから収集し学習事例とする.

続いて, 抽出の手順は以下の通りである.

step1,step2 学習時と同じ.

step3 各文字に対応する素性の束に対し, 学習の手順で構築された分類器を適用してチャンクタグを推定する. 図 4では i 番目の要素のタグを推定するために実線で囲まれた枠内の素性を用いる.i-1 番目,i-2 番目のタグは既に推定されているため解析時に用いることができる. なお上述の手順は, 文末から文頭へ解析 (左向き解析) しているが, 解析方向を逆向きにした右向き解析もある. 日本語固有表現抽出においては左向き解析が有効であることが知られており, 本稿の統計量名の要素の抽出においても左向きで解析を行っている.

7 実験および考察

7.1 実験データ

比較的標準的な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, 再現率で評価を行った. 以下に適合率と再現率について示す.

$$\text{適合率} = \frac{\text{正しくタグ付けされた統計量名の要素の数}}{\text{機械学習によりタグ付けされた統計量名の要素の数}}$$

$$\text{再現率} = \frac{\text{正しくタグ付けされた統計量名の要素の数}}{\text{学習データでタグ付けされた統計量名の要素の数}}$$

チャンキングには SVM に基づくチャンカーである YamCha[YamCha] 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

表 1に各タグの自動抽出の結果である適合率と再現率を示す.

表 1: 各タグの適合率と再現率

	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

7.2.1 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, 再現率の両方が 85 %以上

の精度であったため, 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. 対象に対応する obj に関しては, 属性型の対象と同じタグを使っているため 7.2.2節で考察する.

7.2.2 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り属性に対応する 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. これは動作型の数え方と違い, 属性には様々な表現があり学習が不十分だ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, 動作型と属性型の統計量名の一部であり対象に対応する obj に関しては適合率, 再現率ともに低い精度となった. これは対象となるものが多く学習が不十分だったことが考えられる. また, 動作型の対象と属性型の対象を同じタグでまとめてしまっているが, 属性型の対象は省略されることが多い. 例えば,

> 「自治省は 1997 年度全国人口動態をまとめた. それによると日本の総人口は前年度から 31 万 974 人増えて 1 億 2556 万 8035 人となった.」

という文章では, 「1997 年度の日本の総人口」が統計量名となり, 対象は人であり, 属性が人口となる. したがって対象である人という表現は統計量名では省略されている. またこの人という表現はどの文書にも現れない場合がある. 今後は属性の対象となる表現が省略されることもあるということを考慮し, 動作型と属性型の対象の扱い方について考察していく必要がある.

7.2.3 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. 今回扱った動向情報では「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.

7.2.4 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はある程度の抽出精度と考えられるが, 再現率は全体的に低く, 特に地域を示す locat が低い結果である. これは attr や obj と同様に, 文書に現れる地域の表現が多かったことと, 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. 今後は, 固有表現抽出器の出力と組み合わせることにより, さらなる精度の向上が期待される. 一方で, 期間を示す time に関しては統計を集計した期間以外の情報を抽出してしまっている.

> 「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.」

という文章で「4 月以降の消費税率アップ」は国内販売が落ち込んだ原因と考えられる. このような統計量と関係のない期間などを抽出してしまっている. MuST コーパスでは原因などの情報は del 要素が付与されているが, 学習によって del 要素などのタグを付与することは困難である. そのため, 誤って抽出した期間に関しては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での課題となると考えられる.

7.3 MuST タグとの比較

複雑な統計量を収集して要約や可視化を行うためには, 共通部分と差異の部分を区別する必要があり, 本稿では, 統計量名を要素ごとに分類することを考え, 統計量名の各要素を抽出するためのタグを用意した. 統計量の名前を要素ごとに抽出する手法は, 統計量の名前を文書に現れるそのままの形で抽出する方法と比べ, 抽出精度がどのように変化したかを調べるために比較実験を行う. MuST コーパスで用いられれいる name 要素は, 以下の 3 つの例文に示すように文書に表出する統計量の名前そのものにタグが付与されている.

例文 8 「<name> 乗用車の生産台数 </name> は 473 万台 5374 台で, 前年同期比 12・ 1 %の大幅減少であった.」

例文 9 「日本市場に最も影響のある <name> 中東産ドバイ原油価格 </name> は 2 月から高騰した.」

例文 10 「<name> 総量 </name> は前年同月比 8・ 5 %減の 3632 万ケースだった.」

したがって,name 要素は統計量の名前が文書に現れるそのままの形にタグを付与されたものと考えることができるため,name 要素と提案した統計量名の各要素の主要部分との抽出精度を比較することで, 統計量の名前を要素ごとに抽出する手法と, 統計量の名前を文書に現れるそのままの形で抽出する方法を比較できると考えた. name 要素の抽出方法は 7.1節と同等の手法を用いた. ただし,MuST コーパスでは unit 要素で囲まれている文にしか name 要素が存在しないため,unit 要素で囲まれた文以外は実験では扱わなかった. 結果を表 2に示す.

表 2より, 提案した統計量名の各要素の適合率, 再現率 (obj は除く) は name 要素より良い結果であることがわかる. name 要素は例文 8 では「乗用車の生産台数」, 例文 9 では「中東産ドバイ原油価格」であり, 例文 10 においては「総量」である. このように統計量の名前は様々な表現で文書中に表出するため, うまく学習できず精度が低くなったと考えられる. しかし, 本稿で定義した統計量の名前を各要素ごとに抽出する手法は, 全く違

表 2: MuST コーパスの name との比較

	提案した要素のタグ					MuST
	obj	foot	head	attr	def	name
適合率	76.4	80.1	86.0	80.7	84.7	73.7
再現率	64.4	79.3	85.4	74.6	79.3	74.7

う構造を持つ統計量の名前や,「総量」のように統計量の名前の一部の表記にも対応できるため抽出精度が向上したと考えられる．したがって,統計量名を要素ごとに抽出する方法の有効性は示された． しかし,obj に関しては name 要素より再現率が低い．この原因として,obj は統計をとった対象のみを抽出しているが,name 要素は統計量の名前全てに対してのタグであるため, 対象以外の統計量の名前の部分で抽出精度が上がったため, 表現が多い obj より高くなったと考えられる． また,MuST コーパスの unit 要素は, 統計量に注目している文を限定するためのものである． そのため,name 要素の抽出実験に扱った文は, 全て統計量に関する文であり比較的良い結果が得られたと考えることができる．すなわち, 全ての文を実験に用いた場合は精度は低くなると考えられる．

8 まとめ

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．統計量名は様々な要素から構成されているため, 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.

抽出実験により, 統計量名の構造を分類することで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．また,MuST コーパスで用いられている統計量の名前を表す name 要素と抽出精度を比較をすることで, 統計量名を要素ごとに抽出するほうが精度が高くなることが示された．また, 統計量名を要素ごとに抽出することで, 可視化などを行う際に統計の調査方法が同一のものであるかを判定しやすくなると考えられる.

統計量とは関係のない期間などを抽出してしまっているが, これらは統計量名の抽出タスクの 2 つ目である「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」において除去することができると考えられる．すなわち, 統計量名の抽出には,

- 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

の 2 つの課題が残っている．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] Masayuki Asahara, Yuji Matsumoto. Japanese Named Entity Extraction with Redundant Morphological Analysis. In Proc. HLT-NAACL 2003, 2003.

[Sang 00] E.F.T.K.Sang. Noun Phrase Recognition by System Combination. In Proc. NAACL00, pp,50-55, 2000.

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

[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. 情報処理学会自然言語処理研究会, 2004-NL-164, pp.89-94, 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. 情報処理学会, 自然言語処理研究会報告, 2001-NL-164, 2001.

[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, 自然言語処理研究会報告, 1998-NL-125, 1998.

[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.

[内元 00] 内元清貴, 馬青, 村田真樹, 小作浩美, 内山将夫, 井佐原均. 最大エントロピー法と書き換え規則に基づく日本語固有表現抽出. 自然言語処理, Vol.9, No.1, pp.63-90, 2000.

[山田 02] 山田寛泰, 工藤拓, 松本裕治. Support Vector Machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol.43, No.1, pp.44-53, 2002.

[山本 06] 山本健一, 殿井加代子, 谷岡広樹. タグ付きコーパスを用いた動向情報とその要因の可視化. 言語処理学会第 12 回年次大会ワークショップ, 言語処理と情報可視化の接点, pp.13-16, 2006.

[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, 言語理解とコミュニケーション研究会報告, NLC2005-119, 2006.

付録 統計量名の要素に注釈付けするタグの仕様

- 動作型に関するタグ

  obj 「ビール」「自動車」等, 統計量の値を集計した対象となる部分. ビールの銘柄や自動車の車種名も含まれる.

  foot 対象が受けた動作を表す部分.「出荷数量」における「出荷」等. 対象がした動作を表すのではなく, 対象が何らかの動作を受けることによって生じる統計量の値に対応する統計量の名前の対象が受けた動作を表す表現にこのタグを付与する. すなわち, 対象物が受身となる形の動作.

  head 統計量の数え方.「生産台数」における「台数」等. 動作型の統計量名では head タグは foot タグと対応して現れるが, どちらかが省略される場合もある.

  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 「石油」等, 統計量の値を集計した対象となる部分.

  attr 対象の属性を表す部分.「原油価格」における「価格」等.「人口」も「人」という対象が持つ属性と考えられるのでこのタグが付与する. また「ソニーの利益」などはソニーが様々なものを売ることによって得た利益であり, 様々なものの属性であると考えられるため「利益」もこのタグが付与される.

- 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」「国内総生産」等. 動作型や属性型のような対象物が存在せず, その名前自体が統計量の名前となるもの.

- 「条件」に関するタグ (上記, 統計量の各型に共通)

  time 「10 日」「1998 年」「6 〜 8 月」等, 統計量の値を集計した期間を表す部分. その期間に, 統計量の値が表されている場合や, 統計量の値ではないが, その値の差や比, 順位等の統計量の値の相対値が表されている場合に付与する.

  locat 「全国」「アメリカ」「首都圏」等, 統計量の値を集計した地域を表す部分. time タグと同様に, 統計量の値が表されている場合や, 統計量の値が相対値として表されている場合に付与する.

  agent 統計量の値を集計した対象を発行した会社名や機関名など. 「アサヒビールのビール出荷数量は」という文においての「アサヒビール」や,「ソニーの経常利益は」という文においての「ソニー」に対応する.「松下電器の社長は語った」などのような統計量に関係のない場合はこのタグは付与しない.

  age 「0 〜 15 歳まで」のように統計量の値を集計した対象の年齢. 統計量に関する情報以外にはこのタグは付与しない.

  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. 「完全失業率」の「完全」や「課税出荷数量」の「課税」に対応する. また「自発的に職を失った人の失業率」においての「自発的に職を失った人」にも対応する.

文書中のすべての統計量名には id が付与されており, 各タグにはその要素がどの統計量名に対応しているかを表す id 属性が与えられている. 同じ id を持った要素の集合をまとめることによって一つの統計量名が構成される. 一つの要素が同時に複数の統計量名の要素となっている場合, その要素にはそれぞれの統計量名の id が列挙されて与えられる. 例えば, 5.3 節の図 2に示される例においては, その先頭に<agent id="990000000_1, 990000000_2">というタグが存在するが, これは, このタグで指示される「トヨタ自動車」が990000000_1 ならびに990000000_2 という識別子に対応する統計量名の要素になっていることを示している.

なお, 統計量名の識別子はDOCNO_number という形で表されている. DOCNO は文書番号, number は統計量名が文書中に出現した順に付けた通し番号である.

タグを付与するには, まず文書中に存在する統計量名の要素を全て列挙する. 次にそれぞれがどういう統計量名の要素となっているかを判断して文書中に存在する統計量名を列挙し, 通し番号をつける. そして各要素にタグ (および id) を付与する.