# 小規模な文書集合からの語彙獲得法

足立 貴行　山田 節夫　永田 昌明
日本電信電話株式会社　NTTサイバースペース研究所

## 1. はじめに

現在, Web 上にはニュース記事など様々な情報が存在している. その中から, 最近話題となっている情報を得るのは容易ではない. 例えば, Web 検索を利用して最近話題の情報を得るには, 一般的には次のように行う. まず, 知りたい情報に関連するキーワードを思い浮かべる. 次に Web 検索サイトでキーワード検索する. 最後に検索結果から概要や実際の文書を順番に確認していく. しかし, 検索結果の中には過去の情報や話題性の低い情報が混在するので, 知りたい最近の情報を素早く得られるとは限らない. また, 探したい情報に合ったキーワードを思い浮かべるのが難しいこともある. そこで, 話題性の強い最近の話題語を効率的に得るために, ランキングされた話題語から気になるものを選択するだけで, 利用者の欲しい情報が得られることが望ましい.

次に, 最近の文書集合 (例えば, 2 週間分, 約 2MB) と過去の文書集合 (例えば, 最近 2 週間より前 1 年分, 約 100MB) を用いて話題語について考える. 話題語は, 過去の文書集合と比べて出現確率が高いものだといえる. 話題語の話題性の強さ (話題度) は, 最近と過去の文書集合中の語の抽出確率の差で表すことができる. また, 最近の文書集合においては, 出現頻度の高い一般的な語よりも, 比較的低頻度な語の方が最近出現している可能性が高いと考えられる. ただし, 最近と過去の文書集合の量を比較すると, 最近の文書集合は小規模である.

以上から, 本稿における小規模な文書集合からの語彙獲得は, 過去より最近の文書集合で出現確率が高く, 比較的低頻度な話題語を抽出することである.

小規模な文書集合からの話題語抽出は, 対象となる最近の文書集合から語抽出し, 抽出した語から話題語を選択する処理だと考えられる.

従来, 文書集合中の任意の文字 Ngram について出現頻度などから計算される統計量を基に語抽出する方法[1]が報告されている. これは大規模な文書集合から統計量を基に語抽出することが前提である. 話題語抽出のために従来手法を小規模な文書集合に適用した場合, 信頼性の低い, 低頻度語が多く含まれるので, 精度良く話題語を抽出することは難しい. そこで, 小規模な対象文書集合と, 対象とは別の文書集合をまとめた大規模な文書集合に対して従来手法を適用し, 抽出した語について対象文書

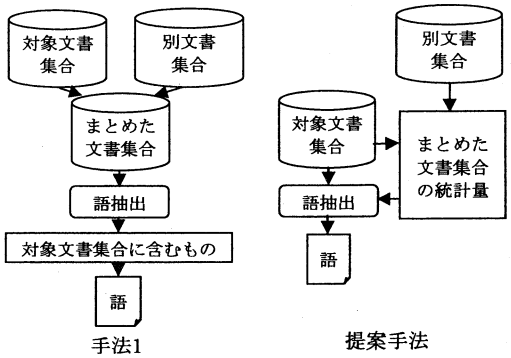


**図1　小規模な対象文書における別文書集合を用いる語抽出手法**

集合に含まれるものに限定する手法(図1の手法1)が考えられる. しかし, まとめた大規模な文書集合を利用しても, 依然精度が低いという問題がある. その原因としては, 語を構成する部分文字列で語とならないものが多く抽出されてしまうためである (例えば, "語彙獲得" の部分文字列 "彙獲得"). そこで, 手法1の語抽出に, 文字列の入れ子となる部分文字列を除外する手法[2]を追加し, まとめた文書集合中の任意の文字列から語候補に絞込むことである程度精度を向上できる.

一方, 話題度順にランキングする場合, 話題語の中でも上位の語が重要視される. そこで, より上位に語でない文字列が抽出されない方が望ましい. しかし, 図1の手法1の語抽出に入れ子文字列の除去処理を追加した手法では, 対象文書集合からでは抽出されなかった語でない文字列が別文書集合も利用することで多く抽出されてしまう (例えば, 対象文書集合では "初生け" しかなかったが, 別文書集合では "初生" があった場合, "初生" が抽出される). そこで, 対象文書集合の任意の文字列から語候補に絞込み, 対象および別の文書集合をまとめた大規模な文書集合から語候補に対する統計量を求めて, 語抽出する手法 (図1の提案手法) の方が話題度順に話題語を抽出する上で適していると考えた.

本稿では, 小規模な最近の文書集合から比較的低頻度な話題語を精度良く抽出し, 話題度順に抽出する際には語でない文字列の数を抑えて抽出する方法を提案する. 以下, 2節で話題語抽出処理の詳細

を述べ，３節で評価実験を行い，４節で考察し，５節でまとめを行う．

## ２．話題語抽出処理の詳細

図１の提案手法に関する処理手順を以下に示す．また，２．１節以降で各処理について説明する．

1. 事前に対象文書集合と別文書集合をまとめた文書集合の統計量を計算しておく
2. 対象文書集合から語抽出を行う
   - 2-1. 任意の文字列作成
   - 2-2. 文字列から語候補への絞込み
   - 2-3. 語候補の統計量を事前に計算した統計量から求め，閾値で絞込み語抽出
   - 2-4. 語の話題度を計算し，話題語選択

### ２．１．語抽出

#### ２．１．１．任意の文字列の作成

対象となる文書集合，もしくは対象と別の文書集合をまとめた文書集合から任意の文字列を作成する．作成には，Yamamoto ら[1]が提案した suffix array を用いた．この方法は任意の文字列を作成した際にその出現頻度やそれを含む文書数も求まる．これらの数値は２．１．３節の統計量の計算に利用する．なお，文書集合中の空白や改行および語となりにくい記号類を特別な文字に置換し，suffix array の作成ではこれらの文字を含まないように変更した．

#### ２．１．２．語候補への絞込み

語候補への絞込みでは，対象文書集合中で頻度２以上の文字列に対し絞り込みを行う．語候補の絞り込みでは，池原ら[2]の方法を用いて入れ子となる文字列を除外した．この方法は，まず，先頭が共通な文字列でその後方文字列を１文字ずつ伸ばした文字列が同じ出現頻度となる場合，最長の文字列を選択する（例：“語彙獲”から１文字伸ばした“語彙獲得”が頻度２以上で最長であり，両文字列とも同頻度の場合，後者を選択）．次に，残った文字列について先頭以外で入れ子となる部分文字列があり，元々同じ位置に出現している場合，入れ子の部分文字列を除外する（例：“語彙獲得”と“彙獲得”が頻度２以上であり，同じ位置に出現する場合，前者を選択）．

その他，以下の処理を施した．

**機能語等を含む文字列の除外**

語と機能語等が連続する文字列を除外する．具体的には，形態素解析を行い，品詞に助詞，接続詞，判定詞，動詞接尾辞を含む場合を除外した．

**文字種・文字数による文字列の除外**

語頭に現れにくい仮名文字は除外する．また，アラビア数字や１文字のものは扱わないことにした．

#### ２．１．３．統計量の計算

２．１．２節で得られた語候補に対して，出現頻度とそれを含む文書数を求め統計量を計算した．出現頻度と統計量の計算は，対象および別の文書集合をまとめた文書集合から求めた．統計量としては RIDF[1]を利用した．RIDF は式(1)のように定義される．

$$RIDF(w) = -\log(df/D) + \log(1-e^{(-tf/D)}) \quad (1)$$

df：文書集合中の語 w を含む文書数
tf：文書集合中の語 w の出現頻度
D ：文書集合中の総文書数

RIDF は，IDF からポアソン分布により推定された IDF を引いたものであり，特定の内容語は同じ文書に何度も出現する傾向にあると仮定して，語らしさの尺度に利用した．df，tf は２．１．１節で述べたように suffix array を用いて求まっている．D は，各文書集合の文書数を求め，足し合わせたものである．最後に RIDF の値がある閾値以上のものを語として抽出する．RIDF の閾値については，３節の評価実験で述べる．

#### ２．１．４．話題語選択

ある語が話題語であるかどうかを最近の文書集合での出現確率が過去の文書集合での出現確率より大きなものとして式(2)を定義した．

$$P\_now(w) > P\_past(w) \quad (2)$$

P_now(w) ：
最近の文書集合での語 w の出現確率
P_past(w)：
過去の文書集合での語 w の出現確率

また，話題語の話題度 T(w)は式(3)とした．

$$T(w) = P\_now(w) - P\_past(w) \quad (3)$$

## ３．評価実験

小規模な対象文書集合から話題語を得るために１節で述べた図１の手法１と提案手法に対して，入れ子となる文字列の除外のあるなしを考慮し，評価した．以下，比較する手法を示す．

a. 提案手法-入れ子除去なし（統計量のみ大規模文書集合を利用，入れ子除去しない）
b. 提案手法-入れ子除去あり（統計量のみ大規模文書集合を利用，入れ子除去する）
c. 手法 1-入れ子除去なし（語候補と統計量で大規模文書集合を利用，入れ子除去しない）
d. 手法１-入れ子除去あり（語候補と統計量で大規模文書集合を利用，入れ子除去する）

### ３．１．評価実験データ

実験に利用したデータを以下に示す.

- 対象文書集合
  京大コーパス[3]が対象としている CD-毎日新聞 95 年版の約半月分 (01/01～01/17) の生データ (記事数約 2300, 文字数約 89 万)
- 別文書集合
  CD-毎日新聞 93, 94 年版 (表 1) から複数準備した. 但し, 特に記載がない場合は 1 年分を利用していることを表す.

表 1 別文書集合

| | 記事数 | 文字数 |
|---|---|---|
| 1 ヶ月分(94/12) | 8362 | 4048268 |
| 2 ヶ月分(94/11-12) | 16601 | 8132581 |
| 3 ヶ月分(94/10-12) | 25309 | 12513287 |
| 6 ヶ月分(94/7-12) | 50416 | 25207221 |
| 1 年分(94) | 101058 | 49630870 |
| 2 年分(93-94) | 170066 | 80556134 |

- 正解データ
  対象文書集合中の任意の名詞相当語句 (名詞, 接頭辞, 未定義語, 接尾辞 (動詞性, 形容詞性を除く)), かつ 2．1．4 節の(2)式を満たす話題語 (語句数約 6.9 万). 名詞相当語句は対象文書集合中の文に対応する京大コーパスのタグ付き文を利用.

### 3．2．評価方法

評価は, 以下の 3 つの項目について行った.

### 3．2．1．入れ子処理の評価

入れ子文字列の除去処理の効果を調べるために, 1 節の図 1 で示した 2 つの手法での入れ子除去あり／なしによって抽出した話題語に対し, 正解データとの適合率 (P), 再現率 (R) から, F 値 (F) を計算し, F 値最大の値を求めて比較した.

P＝獲得した正解話題語数／獲得した話題語数
R＝獲得した正解話題語数／正解データ数
F＝2×P×R／(P＋R)

なお, 出現頻度ごと (2, 3, 4, 5, 6-10, 11-100, 101-1000) に分けて調べた. 話題語抽出では RIDF の閾値ごとに F 値を調べ, F 値最大となる閾値を求めておいた. 用いた RIDF の閾値は-0.2, 0, 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2, 2.5, 3 である. F 値最大となる RIDF の閾値は, 入れ子除去あり (なし) の場合, 出現頻度 1 では RIDF=-0.2(0), 出現頻度 2 では RIDF=0(0.2), 出現頻度 3 以上では RIDF=0.2(0.2)となった.

### 3．2．2．話題度順での話題語抽出評価

1 節の図 1 で示した手法 1 と提案手法に対し, 3．2．1 節で求まった F 値最大となる RIDF の閾値で抽出される話題語について, 2．1．4 節の(3)式の話題度順に並べ, 各順位までの適合率で評価した.

表2 各手法での出現頻度ごとの F 値最大

| 手法 | 提案手法 | | 手法 1 | |
|---|---|---|---|---|
| 入れ子除去 | なし | あり | なし | あり |
| | a | b | c | d |
| 出現頻度 | | | | |
| 1 | — | — | 0.224 | 0.293 |
| 2 | 0.327 | 0.408 | 0.327 | 0.452 |
| 3 | 0.373 | 0.495 | 0.373 | 0.502 |
| 4 | 0.413 | 0.550 | 0.413 | 0.555 |
| 5 | 0.452 | 0.587 | 0.452 | 0.596 |
| 6-10 | 0.489 | 0.615 | 0.489 | 0.627 |
| 11-100 | 0.548 | 0.661 | 0.548 | 0.659 |
| 100-1000 | 0.552 | 0.577 | 0.553 | 0.578 |

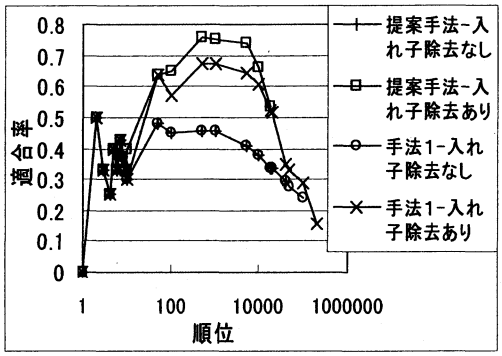


図2 各手法での各順位までの適合率

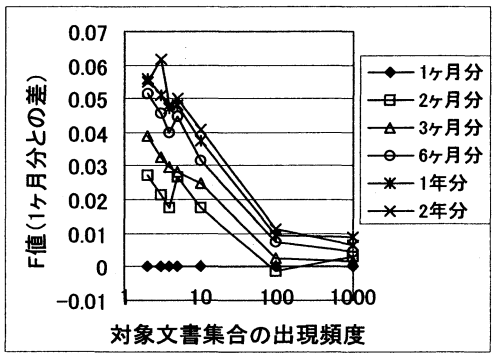


図3 各別文書集合での出現頻度と F 値

### 3．2．3．別文書集合の規模の評価

1 節の図 1 で示した提案手法-入れ子除去ありを例に, 別文書集合の規模による精度の変化を調べた. 別文書集合の規模を 1 ヶ月～2 年まで変え, F 値最大 (もしくはそれと僅差) であった閾値 RIDF=0(出現頻度 2), RIDF=0.2 (出現頻度 3 以上) での F 値 (1 ヶ月分との差) を調べた.

### 3．3．結果

#### 3．3．1．入れ子処理の評価

上記で述べた手法1と提案手法について，入れ子除去ありとなしの場合のF値最大値を表2に示す．各手法とも全ての出現頻度で入れ子除去ありの方が精度は良かった．例えば，提案手法の出現頻度2における入れ子なしとありを比べると，0.327＜0.408となっている．各手法の全出現頻度で入れ子除去をした方が精度は上がっていることから，全出現頻度で入れ子除去は効果があると考える．

#### 3．3．2．話題度順での話題語抽出評価

図2から順位が1～約50位まではどの手法も殆ど変わらないが，100～数万位で提案手法-入れ子除去ありが最も適合率が高くなっている．また，どの手法でも，順位が後になると適合率が低下している．他の手法と比べて，上位からの各順位までの適合率は，提案手法-入れ子除去ありが最も良かったことから，話題度順に話題語を抽出するのに提案手法は適していると考える．

#### 3．3．3．別文書集合の規模の評価

図3からどの出現頻度でも別文書集合を増やすにつれて1ヶ月分とのF値の差は増加している．但し，一般に，別文書集合の規模を増やすと，統計量の信頼性が高くなる効果と，種類数の増加が期待できる効果が考えられる．提案手法-入れ子除去ありの場合は語候補が固定されるので，統計量の信頼性の効果を調べることができる．図3の1000（実際は出現頻度101-1000をまとめたもの）では規模を増やしても影響は少なく安定している．逆に低頻度では規模を増やすと大きく上昇している．但し，約半月分の新聞記事では，ある程度の規模で伸びが鈍化しており，出現頻度を考慮しなければ，1年分あれば十分だと考える．また，規模が増えた際のF値最大となるRIDFは概ねどの出現頻度でも0.2となった．

### 4．考察

#### 4．1．話題語抽出の誤り例

次に，提案手法-入れ子除去ありについて，RIDFの閾値をRIDF=0　(出現頻度2)，RIDF=0.2(出現頻度3以上)とした場合の話題語抽出の誤り例について表記に基づいて調べた結果を以下に示す（括弧内は誤り数）．

- ひらがな1文字＋正解（1729）
  主な原因:形態素解析誤りによる機能語などの未チェック（「のメドゥーサ」）
- カタカナ（747）
  主な原因：正解データとの区切りの違い（「ハクチョウ」），除外すべき文字列の未対処（「オオハ」）
- アルファベット（15）
  主な原因:除外すべき文字列の未対処(「ＵＬ」)
- その他（2154）
  主な原因:除外すべき文字列の未対処(「業制」)，名詞相当語句以外（「参加すべき」）

機能語などの対処については，形態素解析の精度に依存している．この部分は精度の良い形態素解析を利用することが考えられる．正解データとの区切りの違いは，例えば，複合名詞を1つとみなすか複数の語とみなすかというものであり，利用状況によってどちらも正解となりうるものである．今回は特に限定していないので正解とみなして良いともいえる．名詞相当以外については，提案手法が名詞に限定していないためであり，限定するには前処理か後処理に名詞相当語句に絞る必要がある．除外すべき文字列の未対処は，入れ子の処理の改良が必要で今後の検討課題である．

### 5．おわりに

最近の話題語を抽出するために小規模な文書集合から比較的低頻度な語を精度良く抽出し，話題語を話題度順に抽出する際には語でない文字列の数を抑えて抽出する方法を提案した．本手法は，小規模な対象文書中の任意の文字列から，入れ子となる文字列を除去した後，対象および別の文書集合から求めた統計量で絞り込んで話題語抽出を行うものである．評価実験から入れ子除去による精度の向上が確認された．また，話題語を話題度順に抽出する場合，上位から各順位までの適合率を調べた結果，単純に対象と別の文書集合をまとめた文書集合から話題語を抽出し，後で対象文書集合に含むものに絞る手法よりも良い結果が得られた．

今後は，話題語抽出において本手法では未対処となっている語ではない文字列の除外方法について検討したい．

### 参考文献


[1] Mikio Yamamoto and Kenneth W. Church:Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus, Computational Linguistics, Vol. 27, No. 1, pp. 1-30, 2001.

[2]池原悟，白井諭，河岡司:大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法，情報処理学会論文誌，Vol. 36, No. 11, pp. 2584-2596, 1995.

[3]黒橋禎夫，河原大輔:京都大学自然言語処理ツール，自然言語処理研究報告，No. 137-013, 2000.