括弧表現から統計量を用いて有用情報を抽出する手法*

2 R − 3

久光 徹† 丹羽 芳樹†
日立製作所 基礎研究所‡

1. はじめに

新聞記事には，"欧州連合（ＥＵ）"や"朝鮮民主主義人民共和国（北朝鮮）"等の，二つの文字列Ａ，Ｂが括弧により対応付けられた表現"Ａ（Ｂ）"が10行に1個程度の割合で現れ，このようなＡまたはＢの中には多くの重要語が含まれる．以下では，Ａを外側要素，Ｂを内側要素と呼ぶことにする．

本報告では，外側・内側要素の対の出現頻度とパターンマッチのみを用いて，有用な情報を持つ可能性の高い括弧表現を提示し，情報抽出を支援する手法を提案する．

統計的な観点から二つの典型的な型を持つ括弧表現を選出し，適当な言語処理を施すことにより，略称，固有名詞等を大量かつ高精度で獲得できることを示す．

2. 括弧表現の実例

2.1 括弧表現の数

我々が用いた資料（日経新聞1年分）によれば，括弧表現"Ａ（Ｂ）"は，全体で292,799回出現し，その異なり数は177,098であった．このうち2回以上出現したものは，25,421種類であった．

2.2 括弧表現の例と分類

以下に典型的ないくつかの例をあげる：

(2-I) 言い替え(外側要素を内側要素に置き換え可)
- 東京商工会議所（東商）
- 国際通貨基金（ＩＭＦ）
- 朝鮮民主主義人民共和国（北朝鮮）

(2-II) 読み
- 酉（とり）---------- 全体読み
- 森本享（すすむ）--------- 部分読み

(2-III) 補足（外側要素と内側要素は交換できない）
- 種子島空港（中種子町）
- ７月３日（金）

(2-IV) 記事分類・トピック
- 日立製作所（会社人事）
- 林健太郎著（読書）

(2-V) その他（numbering等）
- 理由（１０）

これらの分類は必ずしも網羅的でなく，(2-III)と(2-IV)の間の明確な区別は難しいが，多くの有用情報が括弧表現から抽出できることがわかる．特に注目されるのは，(2-I)と，(2-IV)の一部の要素である．多くの組織名・略称等が(2-I)型括弧表現の中から，多数の企業名・人名が(2-IV)型括弧表現の特定の内側要素に対応する外側要素として得られることである．これは情報抽出の観点から極めて興味深い．

しかし，括弧表現を個別に眺めるだけで有用性を判別することは困難である．例えば，"ＮＫＫ（日本鋼管）"[頻度3]が「言い替え」の括弧であり，"ＮＫＫ（会社人事）"[頻度11]が「記事分類型」の括弧であることを，字種，文字列の包含関係，長さ等を用いたルールだけで判定するのは困難であろう．括弧表現の大域的な出現状況も考慮する必要がある．

3.統計的性質からみた括弧表現の主要型

上記の二つの主要なタイプの括弧表現は，統計的には次のように特徴付けられる：

(3-I) 1対1型（言い替え型）
特定の外側要素と内側要素の対が強く共起する．(2-I)と(2-II)に対応．
例）朝鮮民主主義人民共和国（北朝鮮）
　　酉（とり）

(3-II) 多対1型（分類型）
特定の内側要素に対して，多数の異なる外側要素が共起するもの．(2-IV)に対応．
例）日立製作所（会社人事）

有用な括弧表現抽出のためには，(3-I)型の特徴を持つ括弧表現や(3-II)型の特徴を持つ括弧表現を選出し，言語処理のルールと組み合わせて更に選別を行うことが考えられる．

4. 統計量の利用

4.1 1対1型の括弧表現

1対1型の括弧表現を推定するためには，外側要素Ａと内側要素Ｂの共起の強さを計る必要がある．そのための統計量として，A)頻度，B)相互情報量(MI)[1]，C)χ^2検定，D)「外側要素と内側要素の出現が非独立とする最尤モデルと独立とする最尤モデルの間の尤度比」[2]の4種類を試みた．但し，D)に関して実際は，最尤独立モデルと最尤非独立モデルの対数尤度差の形である次式を用いた：

$$\sum_{i,j} n_{ij}\left\{-\log_2\frac{n_{i\cdot}\times n_{\cdot j}}{N^2}+\log_2\frac{n_{ij}}{N}\right\},$$

$$n_{i\cdot}=n_{i1}+n_{i2},\ n_{\cdot j}=n_{1j}+n_{2j},$$

$$N=\sum_{i,j} n_{ij}.$$

ここで，n_{ij}は以下で与えられる：

---

* Information Extraction from Parenthetical Expressions by Using Statistical Measure

† Hisamitsu, Toru, † Niwa, Yoshiki
‡ Advanced Research Laboratory, Hitachi, Ltd. Hatoyama, Saitama 350-03, Japan

	内側要素がB	内側要素が¬B
外側要素がA	n_{11}	n_{12}
外側要素が¬A	n_{21}	n_{22}

これらを用いて括弧表現をsortした結果，1対1型括弧の推定には尤度比が最も有効であることがわかった．上位100個，500個，1000個までに，(2-I)と(2-II)に属する括弧表現が何個含まれるかを表1に示す．特にB)，C)に関しては，[2]において指摘された低頻度要素の過大評価が，ここでも顕著に見られた．

表1

	尤度比	頻度	χ^2	MI
～100位	90(2)	83(2)	16(0)	0
～500位	418(20)	335(11)	61(1)	1(1)
～1000位	727(46)	554(34)	114(11)	2(1)

*()内の数字は半正解(適切な文字列の削除/追加により正解となる)ものの数

4.2 多対1型の括弧表現

D)によるsortの順序を逆転すれば，典型的な多対1型括弧表現が上位に現れるが，事後エントロピーを利用すればこれらをより鮮明に特徴付けることができる．ここで内側要素Bに対して外側要素$A_1,\ldots,A_m$がそれぞれ$f_1,\ldots,f_m$回共起するとき，Bを固定したときの事後エントロピーE(B)は次式で定義される：

$$E(B) = -\sum_{i=1}^{m} \frac{f_i}{F}\log_2\frac{f_i}{F},\ F = \sum_{j=1}^{m} f_j.$$

E(B)により内側要素Bをsortすると，例えば"会社人事"，"決算数字"等，数千の企業名と共起する内側要素が上位に浮上し，上位語を検討することにより情報抽出上の重要な手掛かりが得られる．表2に，内側要素が漢字列であるものの上位10位を示す：

表2

決算数字;6599;11.30	東京;748;9.23
会社人事;9653;10.90	本社東京;929;9.28
死去;2066;10.82	有価証券含み損;998;9.12
業績修正・配当異動;1332;9.89	仮称;881;8.97
ニューフェース;1033;9.58	読書;521;8.90

*内側要素;頻度;事後エントロピー の順．左列が1～5位，右列が6～10位

5. 統計量と言語処理を併用した有用表現抽出

5.1 1対1型の場合

言い換え型，読み型の括弧表現"A（B）"を抽出する場合，まず尤度比により括弧表現をsortし，1位から，はじめて頻度1の括弧表現が現れる直前までを判別の対象として選択し(異なり数約8,700)，以下の手続きに従って判別を行った：

If　外側または内側要素の字種が，英数字（英字を含む）または片仮名のみ→言い換え
Elseif 内側要素の字種が平仮名のみ→読み
Elseif Aを外側要素としたときの内側要素の集合と，Aを内側要素としたときの外側要素の集合との積集合にBが含まれる→言い替え
Elseif A，Bともに年号→言い替え
Elseif {"頻度によるA（B）の順位" ＞ "尤度比によるA（B）の順位"} ∧ "BはAの弱部分列" → 言い替え（略称）
Elseif {"頻度によるA（B）の順位" ≦ "尤度比によるA（B）の順位} ∧ "BはAの強部分列" → 言い替え（略称）
Else reject

ここで，「BがAの強（弱）部分列」とは，Bの文字全部が順序を保ってAに埋め込まれる（Bの長さが3以上かつその半分以上が順序を保ってAに埋め込まれる）こととする．表4に，上位500個の括弧表現に対する判定精度を挙げる．以下で例えば"正→正"は，目的とする括弧表現を正しく判別した場合，"誤→誤"は，獲得すべきでない括弧表現を正しくrejectした場合で，"正→誤"，"誤→正"(網掛け部分)は誤判別である．尤度比以外の場合は，手続中「尤度比」をそれぞれ対応する統計量に置き換えて上位500個を調べた結果である．尤度比を用いた場合，目的とする括弧表現が上位に集中し，かつ順位の情報が部分列条件の強弱に反映できるため，頻度情報のみを用いた場合に比べ相当高精度となった．

表4

	尤度比	頻度	χ^2	MI
正→正	406(16)	308(6)	61(1)	1(1)
誤→誤	72	151	344	449
正→誤	12(5)	27(5)	0	0
誤→正	10	14	95	50

*()内の数字は，半正解の数

5.2 多対1型の場合

例えば内部要素として"決算数字"と"会社人事"を選んだとき，異なり数3,255個の企業名が外側要素として獲得された．更に表2中の"本社東京"を一般化した"本社＊"を用いると，異なり数13,205個の企業名が外側要素として獲得された．詳細は別稿に譲る．

6. まとめ

新聞記事の括弧表現"A（B）"から有用な情報を抽出するうえで，尤度比とエントロピーを用いて二つの典型的な型の括弧表現を選出し，適当な言語処理を施すことにより，略称，固有名詞等を大量に精度良く獲得できることがわかった．

参考文献

[1] Fano, R. (1961). *Transmission of Information.* MIT Press.

[2] Dunning, T., "Accurate Method for the Statistics of Surprise and Coincidence", *Computational Linguistics*, Vol. 19, Number 1, pp61-74 (1993)