コーパスからの呼応表現自動抽出手法の評価

山本英子[1]　木田敦子[2]　神崎享子[1]　井佐原均[1]

[1]通信総合研究所　　[2]計量計画研究所

eiko@crl.go.jp　akida@ibs.or.jp　kanzaki@crl.go.jp　isahara@crl.go.jp

1. はじめに

日本語において，述語が文の叙法表現の中枢であり，述語の叙法によって文の叙法性が決定される．文の漸進的理解を可能とするには，その文の叙法性を得ることだと考える．また，叙法性に関わりを持つ副詞は陳述副詞と規定されている．陳述副詞の存在は述語の表現を決定付けるため，その文の叙法性を得ることに役立つ．この「陳述副詞によって述語の叙法が決定する」という関係は「ある副詞はある述語表現と呼応する」という呼応関係と見なされる．古語では，係り結びの用法がこの関係を持っていた．現代語において，係助詞と似た役割を果たす副詞が陳述副詞である[10]．これまでに，呼応表現に関する研究として，陳述副詞の記述方法について検討するために 84 の作品から人手によって呼応表現を含む用例を採集し，各用例の頻度と合わせて提示されたものがある[5]．しかし，あらかじめ人手で呼要素も応要素も限定すると，未知の呼応表現を得ることができず，限られた用例しか採集できない．実際，抽出した用例は多いものでも 200 文に足りず，著者自身，採集対象とした資料が少なすぎると明示している．また，人手による作業は，コストが非常に高いという点も問題である．一方，コーパスから自動的に呼応関係を含む用例を抽出すると，コストがかからず多くのバリエーション豊かな呼応表現が得られる．また，コーパスから網羅的に集めた多くの呼応表現は，日本語研究のための基礎資料にも寄与する．そこで，本研究では，より多くの呼応表現を収集するために，コーパスから網羅的に呼応表現を抽出することを目指す．

2. 呼応表現の抽出方法

呼応表現を抽出するには，通常パターンマッチングを行うか，陳述副詞と述語表現が共起しているかどうかを調査する．ここで，「共起」することと「呼応」することとは，平行関係にあることも多いが，原理的に区別すべきかもしれないと指摘されている[5]．これは，共起関係とは同居を表す形式的関係であるのに対して，呼応関係とは結びつきを表す意味的関係であるためである．そこで，「基本的には呼要素と応要素は同じ文中に共起する」と仮定し，大規模なコーパスからさまざまな分野で用いられている尺度で呼応表現の候補を網羅的に抽出する．検討する尺度は，コーパス中の出現頻度情報を用いて，語や文字列の出現状況の類似性を測る尺度である．実験では，

- 共起頻度(co)[8]

$$Co - oc = a$$

- カイ二乗値(chi2)[1,8]

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

- 補完類似度(csm)[12]

$$CSM = \frac{ad-bc}{\sqrt{(a+c)(b+d)}}$$

- ダイス相関係数(dice)[8]

$$Dice = \frac{2a}{2a+b+c}$$

- 対数尤度比(llr)[1]

$$LLR = a\log\frac{an}{(a+b)(a+c)} + b\log\frac{bn}{(a+b)(b+d)} + c\log\frac{cn}{(a+c)(c+d)} + d\log\frac{dn}{(b+d)(c+d)}$$

- 自己相互情報量(pmi)[1,8]

$$PMI = \log\frac{an}{(a+b)(a+c)}$$

- イエーツの補正公式(yates)[1]

$$Yates = \frac{n(|ad-bc|-n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

の 7 つの尺度を呼応表現自動抽出に適用する．そして，網羅性を比較することによって，呼応表現を抽出する問題における適応性を評価する．

3. 実験の概要

3.1. コーパス

実験は新聞記事データを対象とし，毎日新聞記事データ，読売新聞記事データ，日本経済新聞記事データの三種を含む．そのうち，毎日新聞記事データは 1991 年から 2002 年までの 12 年分，読売新聞記事データは 1987 年から 2001 年までの 15 年分，日本経済新聞記事データは 1990 年から 2000 年までの 11 年分である．コーパスの大きさは 4Gbyte，38,875,937 文，1,086,990,614 形態素，年版ごとの形態素種は平均 123,834 である．実験では，このコーパスを用いて下記の工程を経て呼応表現を自動抽出する．

3.2. 対象とする語彙

呼要素となる語は係助詞やある種の副詞と考えられている．しかし，応要素は叙法性を表す部分であるため，特に形式が限定されていない．そのため，応要素を形成する語彙は動詞や助動詞の他に，固有名詞，代名詞，一般名詞を除く名詞や助詞も対象となる．そこで，本研究では，形態素単位の ngram を応要素の候補とし，この ngram と呼要素の候補である副詞と係助詞について，各尺度によって出現状況の類似度(スコア)を求め，呼応表現を抽出する．

3.3. 抽出工程

工程 1 から工程 7 を経てコーパスから呼応表現の自動抽出を行う．

工程 1　新聞記事を一文一行に整形.
工程 2　茶筌による形態素解析.
工程 3　呼要素の候補リストを作成.
　　(副詞と係助詞の抽出．)
工程 4　応要素の存在範囲を特定.
　　(呼要素の候補以降の部分の抽出．)
工程 5　出現状況の数え上げ.
　　(呼要素候補と応要素の存在範囲内の任意の ngram 間を対象．)
工程 6　候補間のスコアを計算.
工程 7　スコアの降順にソート.

4. 評価に用いる正解データ

実験結果を比較・評価するために，認知度が高い陳述副詞「きっと，たぶん，おそらく，決して」に関して正解データを二通りの方法で作成した．一つはコーパスからサンプリングした文に呼応表現タグを人手によって直接付与する方法である[4]．もう一つは実験によってそれぞれの尺度が得た結果の上位をプーリングしたものを文献に明記されている呼応表現[5,10]と人手によってタグ付与されたサンプル文を考慮しながら，正解とするかどうかを判定する方法である．

4.1. 呼応表現タグ付き正解データ

一つ目の正解データは，コーパスから対象とした陳述副詞を含む文をランダムに 1200 件ずつ抽出し，人手により呼応表現タグを付与したデータである[4]．このデータでは，応要素を語尾や動詞，助動詞に限らず，助詞などにつかない無標(unmarked)の形式について 0 要素を応要素として認めている．この考え方は，文献[5]に負っている．このようにタグ付けされた呼応表現をマージすると，呼応表現は各副詞に対して 30～150 件程度である．

4.2. プーリング正解データ

もう一つの正解データは，各尺度によって得た結果の上位 500 件をプーリングしたものである．二つの方法の相違は，前者の方法では作業者に与えられる情報がサンプル文の形態素解析結果だけであるのに対して，後者の方法では前者の方法によって得た正解データを参照しつつ，すべての尺度の抽出傾向が考慮される．この方法で正解データを作成する理由は，コーパスを使う利点であるテキストの量から得られる特徴を残し，かつ尺度に偏らず，呼応表現を判定するためである．このように正解基準をコーパスに依存する形にすることによって，応要素のバリエーションが広がり，呼応表現を網羅的に抽出する目的にあった評価となると考えた．

5. 比較評価

5.1. 精度の評価

精度の評価は，適合率と R-精度を利用する．適合率は上位 N 件を候補表現として精度を測る指標である．R-精度は再現率を考慮した適合率で，R を正解の総数としたとき，上位 R 件を候補表現として，もし候補表現がすべて正解ならば最高値 1 を得る指標である．定義式において，r は候補表現中の正解数である．また，各尺度において同じ値を持つ候補表現については，共起頻度の降順かつ辞書順に並べることによって，順位を統一した．

$$\text{適合率} = \frac{r}{N} \qquad R-\text{精度} = \frac{r}{R}$$

5.1.1. 呼応表現タグ付き正解データによる評価

呼応表現タグ付きデータから抽出した呼応表現に関する適合率を表 1 に示す.

表 1　適合率(上位 500 件)

	Co	Csm	chi2	dice	llr	pmi	yates
きっと	.40	.48	.40	.42	.50	.35	.44
決して	.07	.10	.07	.11	.13	.09	.07
おそらく	.45	.54	.46	.48	.56	.39	.48
たぶん	.44	.46	.42	.42	.47	.39	.48

表 1 から，対数尤度比(llr)が最も高い適合率を示すことがわかる．これは，この正解データがコーパスからのランダムサンプリングによるものなので，必然的に高い共起頻度を持つ呼応表現に偏っているためである．つまり，共起頻度に大きな重きが置かれる尺度は高い適合率を得ると予測できる．このことから，検討した尺度のなかで，llr は共起頻度(co)に最も類似した性質を持つと考えられる．

5.1.2. プーリング正解データによる評価

プーリングした正解データを用いて，比較評価を行う．表 2 に適合率，表 3 に R-精度を示す．それぞれ陳述副詞による結果に対して，適合率と R-精度を計算したものと，その平均である．

表 2　適合率(上位 500 件)

	Co	Csm	chi2	dice	llr	pmi	yates
きっと	.44	.74	.70	.56	.64	.64	.73
決して	.21	.55	.52	.42	.50	.51	.56
おそらく	.33	.56	.56	.54	.51	.46	.59
たぶん	.48	.52	.51	.51	.53	.45	.54
平均	.36	.59	.57	.51	.54	.52	.61

表 3　R-精度

	Co	Csm	chi2	dice	llr	pmi	yates
きっと	.34	.58	.51	.47	.41	.44	.54
決して	.18	.46	.39	.31	.32	.33	.41
おそらく	.29	.51	.47	.46	.38	.36	.50
たぶん	.43	.48	.48	.48	.47	.43	.51
平均	.31	.51	.47	.43	.40	.39	.49

表 2 より，実験において，全体的に適合率が最も高い尺度は yates であった．次に csm，chi2 と続く．これは，期待頻度が低い場合に精度をあげる性質を持つ yates は，他では上位で得にくい共起頻度の低い呼応表現も抽出できたためである．

一方，表 3 より，全体的に R-精度が最も高い尺度は補完類似度であった. 次に yates, chi2 と続く. これは，csm が再現率を重視した場合，共起関係ではなく，呼要素と応要素の出現状況の重なり度合いを測るため，他の尺度に比べ，共起頻度の高低に影響を受けずに呼応表現を抽出できたためである.

また，「決して」に関して，co は他の尺度に比べ，適合率と R-精度のどちらとも非常に低い．これは，「決して」に関する呼応表現の多くが高い共起頻度を持たないためである.

5.2. 尺度間の相関

表 4 にプーリングした正解データに関する順位相関を示す．0.7 以上の相関は太字，0.4 未満の相関は斜体で表す.

表 4　尺度間の順位相関(中央値：.59)

	Co	csm	chi2	dice	llr	Pmi
Csm	.52					
chi2	***.26***	.61				
Dice	.42	.54	.66			
Llr	**.81**	.63	***.38***	.52		
Pmi	***.26***	.51	**.79**	.59	***.33***	
Yates	***.28***	.64	**.93**	.66	.40	**.73**

この表から，最小の相関を持つ尺度対は共起頻度(co)とカイ二乗値(chi2)，co と自己相互情報量(pmi)の 0.26，最大の相関を持つ尺度対は chi2 とイエーツの補正公式(yates)の 0.93 であった. chi2 と yates の相関が高いのは，yates は chi2 を期待頻度が低い場合の精度を上げるために補正されたものであるため，呼応表現がある程度の頻度を持っているのであれば，この結果は当然である.

対数尤度比(llr)は co との相関が最も高く，そのほかとの相関については co の場合より若干高い値を示している．このことからも，5.1 節で考察するように，llr は co に似た振る舞いをすることがわかる．一方，pmi や chi2,yates は co との相関は低く，それら三つの間の相関を観ると，0.7 以上と高いことがわかる．また，中央値 0.59 を基準として観ると，補完類似度(csm)とダイス相関係数(dice)は相関に散らばりがない．以上のことから，7 つの尺度を以下のように分類できる.

- 共起頻度，対数尤度比
- 補完類似度，ダイス相関係数
- カイ二乗値，自己相互情報量，
  イエーツの補正公式

6. 網羅性の向上

本実験では，yates と csm による抽出結果が高い精度を示した．そこで，yates と csm の抽出結果から上位 500 件を統合することを考える．もし統合する抽出結果が同じグループの尺度であるなら，抽出結果の大部分は最も高い精度を持つ尺度の結果に含まれ，精度は大きく向上しないだろう．しかし，順位相関により尺度を分類した結果，yates と csm は別のグループに属するので，大きな精度向上が見込まれる．

表 5 に yates と csm の上位 500 件中のそれぞれの正解数と，重複する正解の数，統合後の正解数を示す．

表 5　統合後の正解数(上位 500 件)

	正解数		正解	統合後
	csm	yates	重複数	正解数
きっと	371	367	261	477
決して	275	280	108	447
おそらく	279	296	163	412
たぶん	260	269	113	416
平均	296	303	161	438

表から大きく正解数が増加することがわかる．重複する正解の数を観ると，yates と csm は抽出結果の上位500件に関して多くの部分が重ならないことがわかる．このことにより，上位 500 件を合わせると，各尺度が単独では抽出できない表現を得ることができる．このように，性質が異なる尺度によって得た結果を統合することにより，抽出結果の網羅性の向上が期待できる．

7. 考察

これまでに，コーパスから知識を獲得するために，ある語間関係を抽出する目的で尺度の性能比較が行われ[6,7,11,13]，また目的にあった尺度選択を支援するツールも提案されている [2]．本研究では，呼応表現を網羅的に抽出する目的において尺度の性能比較を行った．このことにより，本研究は，尺度の新たな適用可能性を検討したと見ることができる．そして，実験結果から，呼応表現を自動抽出する問題に関して，イエーツの補正公式や補完類似度の適用可能性が高いという結果を得た．

8. おわりに

本研究では，コーパスから呼応表現を網羅的に抽出することを目的として，出現状況の類似性を測る尺度によって，自動抽出を試みた．その結果，イエーツの補正公式が最も高い適合率を，補完類似度が最も高い R-精度を得た．この二つは尺度間の相関により，検討した尺度を分類すると，異なる類に属した．このことから，抽出結果を統合することによって，網羅性の向上が期待できることを示した．

謝辞

本研究では，毎日新聞社，読売新聞社，日経新聞社の新聞記事データを使用させて頂きました．また，本稿をまとめるにあたり，通信総合研究所自然言語グループ内山将夫氏に有意義なコメントを頂きました．深く感謝致します．

参考文献

[1] 池田央, 統計ガイドブック, 新曜社, 1989.

[2] 河部恒　柏岡秀紀　田中英輝　松本裕治, 単語類似度の尺度比較支援ツールの作成, 情報処理学会 NL-156-6, pp.39-44, 2003.

[3] Atsuko Kida, Eiko Yamamoto, Kyoko Kanzaki, and Hitoshi Isahara, Extraction and Verification of KO-OU Expressions from Large Corpora. ACL-03 Companion Volume to the Proceedings of the conference, pp.169-172, 2003.

[4] 木田敦子 山本英子 神崎享子 井佐原均, コーパスからの呼応表現自動抽出のための正解データ作成, 自然言語処理学会 NLP2004, D1-3, 2004.

[5] 工藤浩, 叙法副詞の意味と機能－その記述方法をもとめて－ 『国立国語研究所報告 71 研究報告集 3』, 1982.

[6] Lillian Lee, Measures of Distributional Similarity, In 37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp.25-32, 1999.

[7] Dekang Lin, Automatic Retrieval and Clustering of Similar Words, COLING-ACL'98, Vol.2, pp.768-774, 1998.

[8] Christopher, D. Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge MA, 1999.

[9] 益岡隆志 田窪行則, 『基礎日本語文法―改定版―』, くろしお出版, 1992.

[10] 大野晋,『係り結びの研究』, 岩波書店, 1993.

[11] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava, Selecting the Right Interestingness Measure for Association Patterns, SIGKDD'02, pp. , 2002.

[12] 山本英子 梅村恭司, コーパス中の一対多関係を推定する問題における類似尺度, 自然言語処理 Vol.9 No.2 pp.45-75, 2002.

[13] 山本英子 乾裕子 井佐原均, 主観的評価に基づく語間関係の評価尺度の比較, 言語処理学会第9回年次大会, pp.27-30, 2003.