社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－159　(1)
2004／1／13

共起情報に基づく呼応関係自動抽出法の検討

山本英子[1]　木田敦子[2]　神崎享子[1]　井佐原均[1]

概要

本稿では，文を理解する際に役立ち得る呼応関係を頻度情報のみに基づく類似尺度による自動抽出法を検討する．呼応関係とは，一文において，副詞や係助詞が呼びかけとなり，文末がその呼びかけに応答する形式をとる，語間に存在する関係の一つである．この呼応関係を知ることによって，文を文末まで読むまたは聞くことなく，文を理解することが可能となる．呼応関係は呼要素と応要素からなる．本研究では，コーパス中の各文に現れる副詞や係助詞を呼要素の候補，その呼要素の候補より後方に続く形態素を単位とした ngram を応要素の候補とし，いくつかの共起情報に基づく類似尺度を呼応関係の自動抽出に適用した．その結果を比較し，呼応関係抽出法として有効的であった尺度を示す．

Examination of Method for Extracting KOOU Relations based on Co-occurrence Information

Eiko Yamamoto[1]　Atsuko Kida[2]　Kyoko Kanzaki[1] Hitoshi Isahara[1]

Abstract

This paper examines an automatic extraction method by the similar measure only based on frequency information in corpora for a certain concord relation which can be useful in case he understands a sentence. The concord relation that we extract is called KOOU relation, which is particular to Japanese. The KOOU relation is a relation that KO element calls out OU element which responds to its call in a sentence. KO elements are kakari particle or adverb in many cases. OU elements are phrase or string nearby the sentence end. In this study, we consider that all ngrams made of the morphemes which continue the candidate of KO element appearing in each sentence as candidates of OU element, and apply some similar measures based on co-occurrence information to extract KOOU relation. From a comparison of the results, we show the effective measures as KOOU relation extraction method in our experiment.

[1] 独立行政法人 通信総合研究所, Communications Research Laboratory.
[2] 財団法人 計量計画研究所, The Institute of Behavioral Sciences.

1.　はじめに

呼応関係とは，一文中に呼びかける語が出現するならば，呼びかけに応える語が出現するという拘束関係の一つである.「決して行かない」の「決して」と「ない」のように,「先行する一定の語に応じて後ろに特定の形が来る」(『岩波国語辞典』)関係である．これに対して,「共起関係」は,「赤い花」の「赤い」と「花」のように，二つの語が同一文内に出現する関係である. 共起関係には出現順序に制約はないが，呼応関係には出現順序に制約がある. 本稿では以下,「先行する一定の語」を呼要素,「後ろにくる特定の形」を応要素と呼ぶ．古語に存在した係り結びの用法は係助詞と文末の活用形との形態的な呼応関係を持っていた. この用法によって, 係助詞が呼びかけ，文末を決定していたと考えられる．係り結びが消滅した現代語においても，古語の係助詞と似た役割を果たす副詞が存在することが指摘されている[大野 1993]．実際，現代語において「しか〜ない」「決して〜ない」などの呼応関係が存在する．昨今，客観的な基準で作成された実用的な規模の呼応関係データは存在しないけれども，実用的な規模の呼応関係データがあれば対話処理システムに求められる漸進的文理解や文予測のための基礎データとして役立ち，さらに，構文解析の曖昧性解消や係り受け関係を決定するための補助情報としても有効であろう[Kida2003]．そこで，本研究では，客観的かつ実用的な規模の呼応関係データを作成することを目指し，大規模コーパス中から呼応関係を自動抽出することを試みる.

　近年，コーパスから知識を獲得するために，統計や機械学習，データマイニングの分野においてさまざまな尺度が提案されている．これらの多くは語間の関係を獲得することに適用できる．これまでに，目的に合った単語類似度を測る尺度を選択できるように支援するツールが提案されている[河部 2003]．この文献では，各類似尺度の比較評価は行われていないが，単語を単位とする言語知識の獲得や蓄積の延長線上で，単語間の関係を扱おうとする場合，単語同士を比較する何らかの処理が必要となり，適用する目的に最適な尺度の選択が重要であると述べられている．本研究でも尺度選択を重要視する．これまで，さまざまな目的のために尺度の比較が行われている. たとえば，文献[Tan2002]では連想パターンを抽出するために統計や機械学習，データマイニングで提案されている尺度を，文献[Lee1999]では動詞と目的語の対を決定するために確率モデルベースの尺度を，文献[Lin1998]では係り受け関係についてベクトルモデルの尺度を，[山本 2003]ではあるテーマに対する要求の強さを推定する手がかりとなる要求意図表現抽出のためにさまざまな尺度を主観的評価に基づき比較したと報告している．これに対して，本研究では，統計量，確率モデル，ベクトルモデルなどの尺度を語間関係の一つである呼応関係を抽出することに関して比較する.

2.　共起情報に基づく呼応関係抽出法

実験では，大規模なコーパスから漸進的文理解に役立ち得る呼応関係を抽出することを目的とする．そこで，コーパスから呼要素となりうる副詞や係助詞と，応要素を形成しうる動詞や助動詞，固有名詞や代名詞，一般名詞を除く名詞との関係を 2.3 節に示す尺度を用いて抽出する．各尺度は次節に示すパラメータに基づく．各尺度において，スコアが高いほど，二つの要素は関係があると解釈され，スコアの高い順に提示する.

2.1.　共起情報を表すパラメータ

本研究では，呼応要素の相関はコーパス中の出現状況の類似性を表すものとする．そのため，どの尺度も基本的に出現パターンの重なり具合を測定することでスコア(類似度)を得る. 具体的には，コーパスに含まれる文の総数を次元数 n とし，要素が文 i に出現するなら 1，しなければ 0 を置き，出現パターンを二値ベクトル化し，類似度を得る．ここで，各要素の二値パターンを二値 n 次元ベクトル $\mathbf{F} = (f_1,f_2,\ldots,f_i,\ldots, f_n)$, $\mathbf{T} = (t_1,t_2,\ldots,t_i,\ldots, t_n)$とする．ベクトル間の類似度を測るためのパラメータは

　a ：二つの要素がどちらとも出現する文の数
　b ：一方は出現せず，他方は出現する文の数
　c ：他方は出現し，一方は出現しない文の数
　d ：二つの要素がどちらとも出現しない文の数

である．これらは図 1 のように図示できる．ベクトル $\mathbf{F}$ は応要素，ベクトル $\mathbf{T}$ は呼要素に対応する.

		呼要素	
		出現	否出現
応要素	出現	a	c
	否出現	b	d

図 1　共起情報を表すパラメータ

2.2. 検討尺度

本研究では，7 つの尺度を用いて呼応関係の自動抽出を試みた．ここでは，その 7 つの尺度を 2.2 節に示したパラメータを用いた形式で示す．それぞれの尺度に与える引数は二つの要素の出現パターンである．左辺は引数を省略し，各尺度に対応する関数名とした．

- 共起頻度(co)

$$Co-oc=a$$

コーパス中で二つの要素が共起する出現状況を数える関数である．共起頻度が高いほど，二つの要素は関係が深いことを表し，二つの用語間の関係を推定する上で，基となる尺度である[Manning1999]．

- ダイス相関係数(dice)

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

ベクトル空間モデルにおける類似尺度の一つであり，単語間の関係または単語と文書間の関係を推定するために，単語や文書を次元とした多次元空間に配置されたベクトル間の類似度を測る尺度である[Manning1999]．ベクトル空間モデルにおいてもっとも知られている尺度はコサイン関数である．しかしながら，二つのベクトルの大きさが非常に異なる場合，すなわち，出現頻度差が大きい場合，コサイン関数は高い類似度を与えてしまうという問題がある．そこで，この問題を軽減すべく正規化を施した尺度がダイス相関係数である．このような背景から，本研究ではダイス相関係数を検討対象とすることにした．

- カイ二乗値(chi2)

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

共起情報は四つのパラメータで表される．このため，I×J 分割表の特殊な場合にあたる 2×2 表で表現できる．カイ二乗値は標本分布の代表的なカイ二乗分布の値である[池田 1989;Manning1999]．これは独立検定のための尺度なので，二つの要素間に関係がある(独立でない)ならば，高い値を得るという性質を持っている．この性質から，カイ二乗値を検討対象とすることにした．

- イエーツの補正公式(yates)

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

独立検定において標本数が少ない場合，特別な配慮が必要となる．そこで，カイ二乗値に関して，近似の精度をあげるための公式がイエーツの補正公式である[池田 1989]．本研究では，コーパス中に呼応関係を含む文が少ない場合を考慮して，これを検討対象とした．

- 対数尤度比(llr)

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

対数尤度比は二つの確率変数の依存性を表す尺度として一般に知られており，多くの処理に適用されている[池田 1989;Manning 1999]．本研究では，確率変数は各要素，共起確率は a/n,b/n,c/n,d/n, 各要素の出現状況を表す確率は a+b/n, a+c/n, b+d/n, c+d/n に相当し，出現状況をすべて考慮した尺度である．

- 自己相互情報量(pmi)

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

これは対数尤度比の第一項に基づく尺度であり，関連性が高い単語対を発見するために提案された尺度の一つである[Church1990; 北 1999;Manning1999;Rosenfeld1996]．この尺度は一方の単語が出現することによる情報量が他方の単語と共起することを知ることによって増加する情報量を測る尺度である．

- 補完類似度(csm)

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

本来，文字認識の分野で用いられる尺度であり，劣化印刷文字を高い精度で認識できるように経験的に考案された類似尺度である．この尺度は画像パターンをベクトルで表し，劣化印刷文字がテンプレート文字に重なる度合いを測る尺度である[澤木 1995]．この尺度は二つの要素を入れ替えることによって，類似度が異なる非対称性を持つ．この尺度は一対多関係に適用できると報告されている[山本 2002]．

3. 実験の概要

本節では，対象としたコーパスと，抽出工程，抽出された関係の評価方法を示す．

3.1. コーパス

本研究では，新聞記事データをコーパスとして用いた．コーパスは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 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，平均形態素種は 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関係を自動抽出する．

3.2. 抽出工程

図 2 の工程手順で呼応関係を抽出する．

工程 1	各新聞データから一文一行に整形．
工程 2	茶筌で形態素解析を行う．
工程 3	副詞と係助詞を抽出し呼要素の候補リストを作成．
工程 4	呼要素の候補以降の部分を一データとし，集合を一コーパスとする．
工程 5	コーパスから呼要素の候補と応要素の候補となる ngram(1-5 形態素)の出現頻度情報(パラメータ)を得る．
工程 6	呼応関係の候補の頻度情報を用いて，各候補の類似度を計算．
工程 7	類似度の降順に並べ，抽出結果とする．

図 2 呼応関係自動抽出工程

始めに，工程 1 で新聞記事データを一文ずつに切り分け，整形する．これは，呼応関係は一文中で成り立つ閉じた拘束関係であるため，抽出に用いる出現状況の情報を一文ごとに得るためである．工程 2 は文を形態素で分かち書きするためである．工程 3 は品詞情報を利用し，呼要素の候補リストを作成する．このリストと「応要素は呼要素の後方に存在する」という規則に基づき，工程 4 は各文中で呼要素と応要素が存在しうる範囲を限定する．この限定された範囲の文字列を一データとして，出現状況を調査するコーパスを作成する．工程 5 では Suffix Array を用いてすべての文字列の頻度情報を効率的に計算する手法[Yamamoto 2001]を利用し，コーパスから呼要素の候補と応要素の候補となる長さ 1 から 5 までの形態素列(ngram)すべての頻度情報を得る．ただし，対象となる ngram は句読点などの記号は含まない文字列とした．工程 6 はこの頻度情報を用いて，各候補の類似度を求める．最後に類似度が高い順に並べ，その尺度が抽出した呼応関係とする．工程 6 で 3 節に示した 7 つの尺度を適用し，工程 7 で得る結果を比較する．

3.3. 正解データ

対象としてした呼要素「きっと」,「決して」,「おそらく」,「たぶん」が出現する文をランダムに 1000 件抽出し，応要素とみなせる箇所を人手で取り出すことで正解を作成した．たとえば,「名演奏家であれば，きっと，いろいろ気難しいところがあるにちがいない」という文では,「にちがいない」を「きっと」に対する応要素として取り出す．これに対して,「それを乗り越え，別の表現を生み出すパワーが彼らにはきっとある」」のような文では 0 要素を応要素と考える．このように，応要素を語尾や助詞，助動詞に限らず，助詞などがつかない無標(unmarked)の形式については 0 要素を応要素として認める考え方は，文献[工藤 1982]に負う．このように抽出した正解は各呼要素に関して 100～200 件程度である．

4. 実験結果

「きっと」と「決して」について上位に現れた応要素を示す．これらは 5 節で正解と判断される．

- 「きっと」:「う」「だろう」「でしょう」「はず」「かもしれない」「よ」「だ」「くる」「に違いない」「になる」
- 「決して」:「ない」「ではない」「ません」「ず」「わけではなく」「じゃないよ」「なかった」

5. 評価・比較

5.1. 評価対象

本研究では，応要素の候補として対象とする ngram は長さ 1 から 5 までの任意の形態素列であるが，そのうち，際立って頻出する ngram がある．このような高頻度の応要素候補は共起する呼要素が多いため，特徴的な関係といい難い．そこで，高頻度の ngram に関する抽出結果は評価の対象から外した．

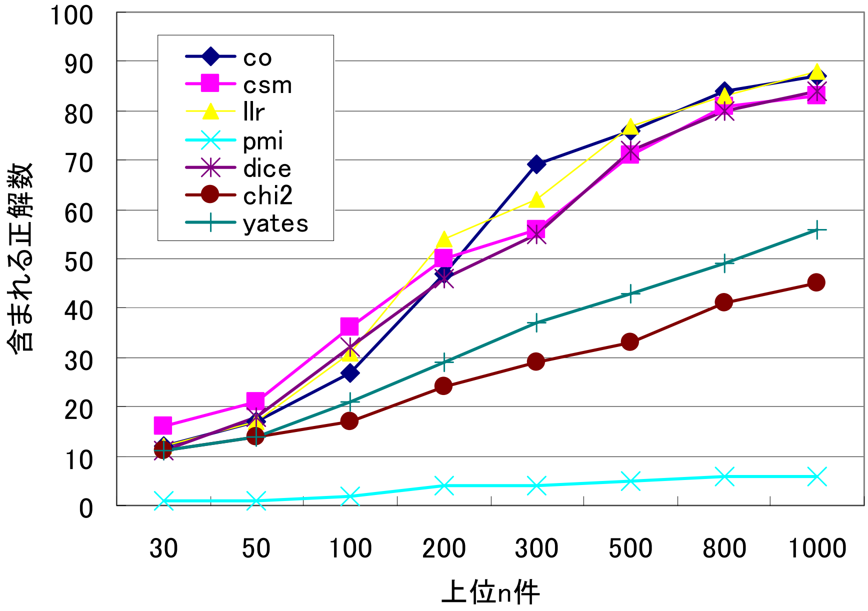

図 3　「決して」に関する関係抽出傾向

5.2.　適用可能性の調査

まず，認知度の高い呼応関係を抽出できるかどうかを調査する．もし認知度の高い呼応関係に高い類似度を保持することができないのであれば，その類似尺度を呼応関係抽出に適用する意味はない．そこで，各尺度の適用可能性を確認した．ここでは，認知度の高い呼応関係として，「決して〜ない」を取り上げる．実験において，この関係は最も共起頻度が高かったため，共起頻度で容易に抽出できる．この関係を検討尺度のうち自己相互情報量を除いては最上位に位置づける．自己相互情報量はこの関係を 5406 位とした．これは，共起頻度は高いが，「ない」の出現頻度が高いため，特異なものとみなさず，高い値を得られなかったためである．このことから，自己相互情報量は本実験の条件下では関係抽出に不向きであると推測される．

5.3.　正解による抽出傾向の分析

図 3 は,「決して」について上位 n 件に何件正解とした関係が含まれるかを調査した結果を示す．同様に「きっと」,「おそらく」,「たぶん」についても調査した．その結果，どの呼要素に関しても，上位で多くの正解を得る尺度は補完類似度であり，上位からある程度の量を見る場合，対数尤度比が多くの正解を得られることがわかった．また，5.2 節で推測したように，その結果は自己相互情報量が本実験では不向きであったことを示している．

5.4.　尺度間の順位相関

表 1 は，正解とした関係について，Kendall の順位相関係数で測った尺度間の順位相関を示す．表 1 から，全体的に対数尤度比はもっとも共起頻度に似た振舞いをすることがわかる．一方，自己相互情報量はどの尺度とも相関が低いことがわかる．これは 5.2 節と 5.3 節と同様，本実験には適していなかったことを示唆する．また，カイ二乗値とイエーツの補正公式は相関が高い．これは,「きっと」に関して正解とした関係にごく稀にしか共起しない関係が含まれていなかったためである．これに対し,「きっと」以外の例では，ごく稀にしか共起しない関係がある程度含まれていたので，際立って相関が高いということはなかった．

表 1　「きっと」に関する尺度間の順位相関

	dice	chi2	yates	llr	pmi	csm
co	.576	.445	.451	.805	.071	.522
dice		.429	.461	.544	.054	.436
chi2			.775	.418	.054	.347
yates				.438	.013	.351
llr					.049	.499
pmi						.054

6.　考察

5 節の比較から，呼応関係抽出法として，抽出の正確さ重視で呼応関係を得るためには，補完類似度が有効であり，量を重視する場合には，対数尤度比が有効であると考察する．また，本実験の条件下における呼応関係抽出法としては，自己相互情報量は不

向きであることがわかった．

抽出された結果には呼応関係と判断できるもののほかに，呼要素「きっと」に対する応要素として「喜んでくれる」，同様に呼要素「決して」に対して「あきらめない」や「楽」，「平たん」，「おそらく」と「たぶん」は「重要」や「必要」などよく目にする表現が比較的上位で抽出されていた．また，「おそらく」と「たぶん」は「将来」や「当時」，「初めて」など時に関する副詞化可能な名詞や副詞との関係が目立った．これらは呼応関係ではないが，有用な情報になり得ると考えられる．

7. まとめ

本稿では，文を理解する際に役立ち得る呼応関係を頻度情報のみに基づく類似尺度による自動抽出方法を検討した．具体的には，コーパス中の各文に現れる副詞や係助詞を呼要素の候補，その呼要素の候補より後方に続く形態素を単位とした ngram を応要素の候補とし，いくつかの共起情報に基づく類似尺度を呼応関係の自動抽出に適用し，結果を比較した．その比較により，本実験において，対数尤度比と補完類似度が呼応関係抽出法として有効的な尺度であると考察した. 今回は, 一つの文献の考え方を基に, 主観的な正解データを作成したため，結果比較による尺度の優越を示すにとどまった．複数の考え方による網羅的かつ客観的な評価は今後の課題とする．

謝辞

通信総合研究所自然言語グループ内山将夫氏に深く感謝いたします.

参考文献

[Church1990] Kenneth W. Church and Patrick Hanks, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, 16(1), pp.22-29, 1990.

[澤木 1995] 澤木美奈子 萩田紀博, 補完類似度による劣化印刷文字認識，95-PRU-14，pp.101-108, 1995.

[池田 1989] 池田央，統計ガイドブック，新曜社, 1989.

[河部 2003] 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.

[Kida2003] Atsuko Kida, Eiko Yamamoto, Kyoko Kanzaki, and Hitoshi Isahara, Extraction and Verification of KO-OU Expressions from Large Corpora, ACL2003 Companion Volume to the Proceedings of the conference, pp.169-172, 2003.

[北 1999] 北研二, 言語と計算 4 確率的言語モデル, 東京大学出版会, 1999.

[工藤 1982] 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.

[Lee1999] Lillian Lee, Measures of Distributional Similarity, In 37th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp.25-32, 1999.

[Lin1998] Dekang Lin, Automatic Retrieval and Clustering of Similar Words, COLING-ACL'98, Vol.2, pp.768-774, 1998.

[Manning1999] Christopher D. Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge MA, 1999.

[松本 2002] 松本裕治 北内啓 山下達雄 平野善隆 松田寛 高岡一馬 浅原正幸, 形態素解析システム「茶筌」Version 2.2.9, 2002.

[大野 1993] 大野晋，係り結びの研究，岩波書店, 1993.

[Rosenfeld1996] Ronald Rosenfeld, A Maximum Entropy Approach to Adaptive Statistical Language Modeling, Computer Speech and Language, 10(13), pp.187-228, 1996.

[Tan2002] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava, Selecting the Right Interestingness Measure for Association Patterns, SIGKDD'02, 2002.

[山本2003] 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.

[山本 2002] 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.

[Yamamoto2001] 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.