E-021

統計指標を用いた日本語コーパスからのコロケーション情報抽出と精度評価

Extracting Collocational Information from Japanese Corpus Using Statistical Indicator

園田 匠 † 島田 諭 ‡ 三浦 孝夫 †
Takumi SONODA Satoshi SHIMADA Takao MIURA

1 まえがき

テキストの翻訳や校正では，特定の語の共起による「習慣的表現」「人間らしい自然な表現」であるコロケーションが重要である．特に，自動翻訳や自動校正などのテキスト処理において高い精度を得るためには，コロケーションの抽出精度を高めることが必要となる．これまで英語のコロケーションについては，言語学をはじめとする多方面からの研究が行われ，工学的な応用も実現してきているが，日本語のコロケーションについて工学的に扱った研究は少ない．

Stubbs [1] は，英語のコロケーションを 4 種に分類しており，日本語においても適用できると考えられる．

コロケーション…語の機械的共起

連辞的結合…語と文法，形態素の共起

優先的意味選択…語と特定意味領域語群の共起

談話的韻律…語の意味，文法を超えた共起

例えば「事件，起こる」「問題，起こる」はコロケーション,「名詞，起こる」は連辞的結合,「何かしらの事柄，起こる」は優先的意味選択,「匙，なげる」は談話的韻律である．これらの分類に属さない，習慣的でない偶発的な共起や,「が，は，を，に」のような語自体に直接意味をもたない形態素を含む共起はコロケーションとはみなされない．

日本語に対応するには屈折語，膠着語という言語の特性の違いがある．屈折語は英語の特性であり，語順により語の格が変化するのに対し，膠着語は日本語の特性であり，例えば「首相“ が ”アメリカ“ に ”訪れる」,「アメリカ“ に ”首相“ が ”訪れる」のように意味が語順でなく格助詞に依存して決定する．このため日本語コロケーションでは語順を考慮しないが，連鎖共起により語の依存性を検討する．

本研究では，拡張 n-Xgram および日本語 POS フィルタを用いてコロケーションを抽出する．また共起頻度，相互情報量，ダイス係数，T スコア，対数尤度比，および連鎖共起を用いてその精度を評価する．2 章では日本語コロケーションの特性を考慮した抽出手法を提案する．3 章では実験を行い，その抽出精度を検討する．4 章にて考察を行い，5 章で結びとする．

†法政大学大学院 工学研究科
‡法政大学マイクロ・ナノテクノロジー研究センター

2 n-Xgram による頻出共起語抽出

この節では n-Xgram，日本語 POS フィルタによる頻出共起語の抽出についてその実現方法を述べる．

2.1 拡張 $n-gram(n-Xgram)$

コロケーション抽出の手法として，我々は自立語を最小単位とした $n-gram$(以下 $n-Xgram$) を提案している [2]．一つの文章に含まれる自立語の連続する語句の数に上限を与え，人工的に文章とみなしてその範囲内での共起を抽出する．表 1 に $n-Xgram$ の例を示す．

例 {すもも も もも も もも の うち}

n	n-Xgram
2	{ すもも, もも }{ もも, もも }{ もも, うち }
3	{ すもも, もも, もも }{ もも, もも, うち }
4	{ すもも, もも, もも, うち }

表 1: n-Xgram

$sentence-Xgram$ (以下 $s-gram$) とは，一つの文章に含まれる全ての自立語の並びを意味する．連続する語句が多くなるほど様々な語の影響を有する特徴的な共起が生じると考えられる．なお，得られた $n-Xgram$ の頻度は n で割る必要がある．例えば $2-Xgram$ にて「すもも，もも」と「もも，もも」を比べると「もも」が重複しているが，連続する語句の先頭と末尾の語句を例外として重複により，語句の重みが変わるためである．

2.2 POS フィルタ

POS(Part Of Speech) フィルタとは，形態素列 (以降”品詞パターン”と呼ぶ) を与え，文書に出現するコロケーションに対して適合する並びだけを出力するものである．表 2 に英語における POS フィルタの例を示す．ここで A は形容詞 Adjective， N は名詞 Noun，P は前置詞 Preposition を示す．

英語では "A N"， "N N"， "A A N"， "N A N"，"N N N"， "N P N" を品詞パターンとして与えているが，日本語に適用するわけにはいかない．コロケーションは一方の語に依存する語句の組合わせとされることが多い．また日本語コロケーションは自立語の共

A N	linear function
N A N	mean squared error
N P N	degrees of freedom

表 2: POS filtering

起である．これらのことから我々は日本語に対応した POS フィルタを表 3 のように提案している [2].

{ 名詞, 形容詞, 副詞 }.* + 動詞
胸 + 張る
しっかり + する
方針 + 明らか + する

表 3: 日本語 POS フィルタ

膠着語特有の語順の自由さに対応するため，日本語を扱う本研究では語順は考慮しない．固有表現についてはコーパスの時系列データに左右されやすい．また例えば「イチロー」が「松井」より「打つ」という語に共起しやすいといった情報は前項での優先的意味選択型にあるように不要な検討であり，談話的韻律に影響しづらい．このため表 4 のように抽象化する．

固有表現	ラベル	例
人名	PER	イチロー 安倍晋三 ビルゲイツ
組織名	ORG	日本テレビ 東京電力 Microsoft
地名	LOC	東京都小金井市梶野町

表 4: 固有表現の抽象化

ここで,「東京都, 小金井市, 梶野町」は「LOC, LOC, LOC」となるが，同じ抽象化ラベルが続いた場合は「LOC」が 1 回出現したものとみなす．また複合動詞, 複合形容詞を構成する日本語特性を考慮しサ変接続の名詞に「する」が後続した場合は連結して品詞を動詞に，ナイ形容詞語幹の名詞に「ない」が後続した場合は連結して品詞を形容詞とする．

2.3 頻出共起語の抽出

高頻度に共起する頻出共起語語を抽出するため，`FP(Frequent Pattern)-tree` を用いる．これは，従来の頻出アイテムセット発見手法である `Apriori` を改良し，逆単調性を活かした高速探索アルゴリズムである [4] ．全てのトランザクションデータを木構造として圧縮し，候補パターンを生成することなく頻出アイテムセットを見つけることができる．

3 実験

この章では前項で述べた 3 つの手法を用いて抽出した共起語に対し，各指標におけるコロケーションの抽出精度を検討する．

3.1 n-Xgram による共起語抽出

読売新聞記事データ集 2007 から 1 月から 6 月分の半年分，2,407,601 文を用い，`MeCab` [1] にて形態素解析を行い，抽象化処理を行う．提案手法の各パラメタは，$n-Xgram$ の n を $2 \sim 5$, $sentence$ の 5 パターン , $\mathrm{FP-Tree}$ のしきい値を 0. 01 として実験を行う．

3.2 統計指標

本研究では共起頻度と，語彙研究にて代表的な指標である [5] 相互情報量，ダイス係数，T スコア，対数尤度比を用いる．

3.2.1 共起頻度

共起頻度 ($Co-OccurencesFrequent$) とは総文章数 N に対する共起の出現回数 n の相対比であり，次のように定義する．

$$freq(n, N) = \frac{n}{N} \times 100$$

高い値を示せば頻度が高く，習慣的に使われる共起であることを意味する．例えば「ご飯, 食べる」という共起が「飲み物，飲む」という共起より頻出なとき,「ご飯，食べる」は強い関係をもつと期待できる．

3.2.2 相互情報量

相互情報量 ($MutualInformation$) は情報検索において，二つの確率変数の相互依存性を意味する指標である．共起する語 W_1，W_2 の出現回数を n_1，n_2，共起回数を n_{12}，総文章数を N とすると，次のように定義される．

$$MI(W_1, W_2) = log_2 \frac{n_{12} \times N}{n_1 \times n_2}$$

本研究においては共起の関連強度を意味する．このスコアが高ければ W_1 は W_2 に，W_2 は W_1 に依存して共起する特徴を示すが，極端に低頻度な語に対し高いスコアを付与してしまう欠点がある．

3.2.3 ダイス係数

ダイス係数 ($DiceCofficient$) は，共起する語 W_1, W_2 の出現回数を n_1，n_2，共起回数を n_{12}，とし，次のように定義される．

$$DC(W_1, W_2) = 2 \times \frac{n_{12}}{n_1 \times n_2}$$

相互情報量に似た指標であるが，総文章数を考慮せずに語の生起、共起回数のみで強度を測る指標である．

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

3.2.4 T スコア

T スコア ($TScore$) は情報検索において，母集合と確率変数の相対的評価を意味する指標である．共起する語 W_1，W_2 の出現回数を n_1，n_2，共起回数を n_{12}，総文章数を N とすると，次のように定義される．

$$TS(W_1, W_2) = (n_{12} - \frac{n_1 \times n_2}{N}) \div \sqrt{n_{12}}$$

本研究では総文章数を考慮した二語の出現比率を意味する．語の生起，共起頻度を大きく重視した算出式であり，頻度の高い共起が高いスコアをもち，頻繁に用いられる，習慣的な共起の評価に優れている．

3.2.5 対数尤度比

対数尤度比 ($Log - LikelihoodRatio$) はある前提条件に従い得られる結果に対し，逆に得られる結果から前提条件を推測する「尤もらしさ」を意味する指標である．スコアの算出に当たり，語 W_1 と語 W_2 の生起，共起情報を下記表 5 に示す．

	W_2	$\neg W_2$	Total
W_1	A	B	C
$\neg W_1$	D	E	F
Total	G	H	I

表 5: 生起, 共起情報

対数尤度比を導く式は，次のように定義される．

$$LLR = 2 \times \sum(\log e(\text{実測値}) - \log e(\text{期待値}))$$

これを表 5 の A $\sim$ I の各値の対数をとり $a \sim i$ とし，次のように変形ができる．

$$LLR = 2 \times \{(a+b+d+e+i) - (c+f+g+h)\}$$

本研究においては，例えば「事件」の出現とともに「起こる」が共起したとき，逆に「起こる」の出現により「事件」が共起することをどれだけ期待できるか推測する．

3.3 連鎖共起

連鎖共起 ($ChainCo - Occurences$) は語の結びつきの強さを意味し，他の指標と違い非対称性である．語 W_1，W_2 の生起回数を $n_1 n_2$，共起回数を n_{12} とするとき,，W_1 の W_2 に対する連鎖共起のスコアは次のように定義する．

$$P(W_1|W_2) = \frac{n_{12}}{n_2}$$

このスコアが高ければ W_1 が W_2 に依存して共起するという特徴を示し，コロケーションとしての価値を持つと考えられる．例えば「噛むのは犬」であっても「犬は噛む以外に吠えもする」．その非対称性から他の指標との比較がしづらいため，本研究では独立に評価を行う．

3.4 評価方法

各指標での結果に対しコロケーションであるかの正誤判定し，精度を評価する．判定にはコロケーション辞典 [6] とオンライン辞書サービス Weblio [2] を参考に，人手で正解データを作成した．

3.5 実験結果

3.5.1 n-Xgram による共起語抽出

頻出共起語の総抽出数を頻度，n-Xgram ごとに表 6 に示す．

	~ 0.1	0.1$\sim$ 0.07	0.07$\sim$ 0.04	0.04$\sim$ 0.01
2-Xgram	17	14	84	876
3-Xgram	22	20	98	1,241
4-Xgram	23	25	118	1,515
5-Xgram	26	33	132	1,715
s-Xgram	225	222	870	6,346

表 6: 頻出共起数

頻出共起の多くは頻度 0.04%以下に多く存在することを示している．
頻出共起の語数の遷移を表 7 に示す．

	2 語	3 語	4 語	5 語以上
2-Xgram	991	-	-	-
3-Xgram	1,292	90	-	-
4-Xgram	1,503	165	13	-
5-Xgram	1,728	165	13	0
s-Xgram	7,485	165	13	0

表 7: 頻出共起語数

ほとんどは 2 語の共起であり，5 語以上の頻出共起は存在しないことを示している．

3.5.2 統計指標による抽出精度

各統計指標でのスコアを高い順にランキングし，高い共起 50 件を TOP50，各スコアの平均から前後 25 件を MID50 とする．これらの精度を図 1，2 に示す．
共起頻度は MID50，その他は TOP50 の精度が 58.0% と高くなることを示し，その他は TOP50 の精度が高い傾向を示す．また最良はダイス係数 TOP50，2-Xgram での 88.0

3.5.3 連鎖共起による抽出精度

連鎖共起のスコアをランキングし，その TOP50 件の精度を表 8 に示す．
2-Xgram のとき **78.0**%と最も高い抽出精度となり，連続する語句の数の上限が増すことでその精度が落ちていくことを示している．

[2] http://www.weblio.jp/

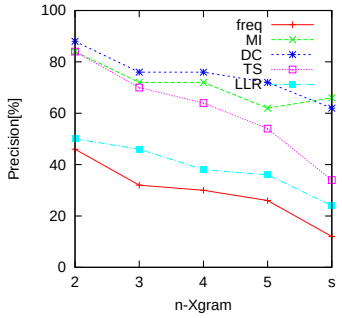

図 1: 統計指標 TOP50 精度

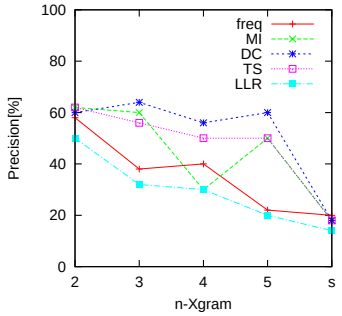

図 2: 統計指標 MID50 精度

n-Xgram	精度
2-Xgram	**78.0**%
3-Xgram	46.0%
4-Xgram	32.0%
5-Xgram	32.0%
s-Xgram	14.0%

表 8: 連鎖共起 TOP50 精度

4 考察

4.1 n-Xgram による抽出精度

n-Xgram による連続する語句の数の制限について，共起頻度，連鎖共起，対数尤度比は上限を増やすごとに精度が低下していることがわかる．その他の指標では上限を増やしても精度はほぼ横ばいで，向上することはなかった．また 5-Xgram 以降は 5 つ以上の語の共起が抽出可能であるが，そのような頻出共起は抽出されなかった．以上より連続する語句の数は 4-Xgram 以内が適しており，それ以降は的確なコロケーション抽出の妨げになることが分かった．

4.2 統計指標での抽出精度

前項の n-Xgram の評価を踏まえ 4-Xgram までを対象とし，TOP50，MID50 での具体例とその正誤を表 9 ～ 18 に示す．*TOP*50 は各指標のスコアを高い順にランキングし，高い共起 50 件を，*MID*50 はスコアの平均付近の共起 50 件を意味する．

共起	freq[%]	正誤	共起	freq[%]	正誤
ORG+発表する	0.020	×	LOC+戻る	0.020	×
LOC+学ぶ	0.020	×	協力+得る	0.020	○
運転+帯びる	0.020	×	LOC+使う	0.020	×
見込み+なる	0.020	○	LOC+含む	0.020	×
姿+見せる	0.020	○	背景+ある	0.020	×
責任+ある	0.020	○	PER+喜ぶ	0.020	○
処分+受ける	0.020	×	LOC+迎える	0.020	○
支持+訴える	0.020	×	ORG+目指す	0.020	×
争点+なる	0.020	○	問題+受ける	0.020	×
しっかり+する	0.020	○	LOC+除く	0.020	×
調査+よる	0.020	×	PER+述べる	0.020	○
間もなく+死亡する	0.020	○	容疑+否認する	0.020	○
指摘+ある	0.020	×	会長+務める	0.020	○
ORG+向ける	0.020	○	現金+奪う	0.020	○
調査+する	0.020	○	PER+訴える	0.020	○
理解+深める	0.020	○	期待+寄せる	0.020	○
人+なる	0.020	×	返還+求める	0.020	○
良い+なる	0.020	○	ない+思う	0.020	×
意見+聞く	0.020	○	納得+できる	0.020	○
はっきり+する	0.020	○	手術+受ける	0.020	○
事件+受ける	0.020	×	LOC+向かう	0.020	○
委員+開く	0.020	×	頭+下げる	0.020	○
気+する	0.020	×	確保+できる	0.020	×
楽しみ+する	0.020	○	原因+なる	0.020	○
大事+する	0.020	○	PER+知る	0.020	×

表 9: 共起頻度 2-Xgram,MID50

共起	頻度 [%]	正誤	共起	頻度 [%]	正誤
LOC+する	0.410	×	処分+する	0.090	○
LOC+ある	0.360	×	容疑+よる	0.090	×
PER+話す	0.350	○	PER+受ける	0.090	×
LOC+なる	0.320	×	ORG+ある	0.090	×
PER+する	0.310	×	PER+務める	0.080	○
LOC+開く	0.280	×	PER+開く	0.080	×
LOC+行う	0.250	×	記者+会見する	0.080	○
PER+よる	0.240	×	PER+語る	0.080	○
明らか+する	0.230	○	ない+する	0.080	×
LOC+よる	0.220	×	原因+調べる	0.080	○
可能+ある	0.200	×	LOC+始まる	0.080	×
PER+なる	0.180	×	人+する	0.080	×
容疑+逮捕する	0.160	×	ORG+よる	0.080	×
PER+ある	0.160	×	調べ+容疑+PER+よる	0.080	×
必要+ある	0.150	○	現行+逮捕する	0.080	×
ORG+する	0.140	×	明らか+なる	0.080	○
LOC+受ける	0.130	×	力+入れる	0.070	○
調べ+よる	0.120	×	LOC+見る	0.070	×
疑い+逮捕する	0.110	×	調べ+PER+よる	0.070	×
容疑+PER+よる	0.100	×	LOC+求める	0.070	×
対象+する	0.100	○	LOC+目指す	0.070	○
ORG+なる	0.100	×	被告+PER+よる	0.070	×
人+いる	0.100	○	恐れ+ある	0.070	○
LOC+訪れる	0.090	○	PER+述べる	0.060	○
PER+行う	0.090	×	ORG+行う	0.060	×

表 10: 共起頻度 4-Xgram,TOP50

共起	MI	正誤	共起	MI	正誤
酒気+帯びる	10.3	○	回+振る	7.54	×
骨+折る	9.44	○	ウイルス+検出する	7.53	○
耳+傾ける	9.23	○	酒+飲む	7.52	○
体調+崩す	9.13	○	うまい+いく	7.52	○
握手+交わす	8.83	○	口座+振り込む	7.46	○
震度+観測する	8.75	○	苗+植える	7.45	○
熱戦+繰り広げる	8.63	○	健闘+誓う	7.42	○
球+振る	8.62	○	詐欺+振り込める	7.30	×
工夫+凝らす	8.44	○	軽傷+負う	7.20	○
アルコール+検出する	8.43	○	重傷+負う	7.03	○
１１０番+通報する	8.35	○	懲役+求刑する	6.90	×
余罪+追及する	8.24	○	胸+張る	6.87	○
カギ+握る	8.22	○	けが+負う	6.85	○
汗+流す	8.21	○	協定+結ぶ	6.84	○
重傷+折る	8.08	×	納得+いく	6.78	○
定年+退職する	8.03	○	失点+抑える	6.76	○
傷+探る	8.02	○	足+運ぶ	6.74	○
ガラス+割る	7.95	○	役割+果たす	6.73	○
全力+尽くす	7.93	○	理解+深める	6.64	○
基金+取り崩す	7.69	×	文化財+指定する	6.59	×
雨+降る	7.66	○	署員+駆けつける	6.59	×
書類+送検する	7.64	○	スタート+切る	6.58	○
正式+決定する	7.57	○	出馬+表明する	6.53	○
支払い+命じる	7.57	○	行方+追う	6.49	○
首+絞める	7.56	○	議案+提案する	6.46	×

表 11: 相互情報量 2-Xgram,TOP50

共起頻度による抽出精度は MID50 での精度が高くなる傾向を示した．具体的な傾向としては動詞は構文上用いられ，語自体に自立した意味をもたない軽動詞になりがちな「する，ある，なる」が，名詞は固有表現を抽象化した「PER，ORG，LOC」が含まれる共起が多くを占めた．高頻度な共起は常識的であり，コロ

共起	MI	正誤	共起	MI	正誤
調査＋始める	0.228	○	問題＋考える	0.002	○
PER＋{ 立候補, 表明する }	0.223	×	選挙＋立候補する	-0.008	○
人＋増える	0.222	○	大会＋目指す	-0.038	○
イベント＋開く	0.213	○	記者＋開く	-0.047	○
気象台＋よる	0.208	×	容疑＋送検する	-0.049	×
ORG＋{LOC, よる }	0.198	×	試合＋行う	-0.050	×
よく＋見る	0.192	×	PER＋{ 強い, 打つ }	-0.052	×
調査＋よる	0.192	×	市長＋話す	-0.055	○
予定＋開く	0.186	×	報告＋受ける	-0.058	○
臨時＋開く	0.180	×	LOC＋{ 調べ, よる }	-0.061	×
PER＋{ 昨年, よる }	0.172	×	教授＋話す	-0.069	×
命令＋受ける	0.158	○	容疑＋供述する	-0.092	○
作品＋見る	0.136	○	市＋受ける	-0.109	×
ホール＋開く	0.103	×	同日＋よる	-0.129	×
活動＋行う	0.093	○	講座＋開く	-0.132	×
LOC＋{ORG, 開く }	0.088	×	利用＋できる	-0.142	×
治療＋受ける	0.080	○	同社＋よる	-0.159	×
問題＋取り組む	0.075	○	3月＋よる	-0.168	×
容疑＋否認する	0.057	○	問題＋巡る	-0.172	○
笑顔＋話す	0.036	○	問題＋解決する	-0.181	○
人＋楽しむ	0.035	○	人＋来る	-0.191	○
文化＋開く	0.023	×	教委＋よる	-0.199	×
問題＋めぐる	0.014	○	PER＋{LOC, 務める }	-0.211	×
相談＋受ける	0.005	○	男性＋話す	-0.234	○
LOC ＋ 大会＋開く	0.004	×	協議＋開く	-0.241	○

表 12: 相互情報量 4-Xgram,MID50

共起	DC	正誤	共起	DC	正誤
よく＋わかる	0.028	○	支援＋受ける	0.026	×
現職＋目指す	0.028	○	大会＋開く	0.026	○
声＋求める	0.028	○	事件＋調べる	0.026	○
指導＋受ける	0.028	○	情報＋提供する	0.025	○
意見＋出す	0.027	×	コメント＋出す	0.025	×
人＋訪れる	0.027	○	県内＋よる	0.025	×
意見＋述べる	0.027	○	理解＋求める	0.025	○
実現＋向ける	0.027	○	委員＋開く	0.025	×
会合＋開く	0.027	×	見方＋示す	0.025	○
作業＋進める	0.027	○	決意＋述べる	0.025	○
調査＋行う	0.027	○	事故＋死亡する	0.025	×
責任＋認める	0.027	○	協議＋めぐる	0.025	×
整備＋進める	0.027	○	意欲＋示す	0.025	○
人＋見る	0.027	×	意見＋出る	0.024	○
候補＋擁立する	0.027	×	中心＋{LOC, する }	0.024	×
積極＋取り組む	0.026	×	容疑＋{ 疑い, 逮捕する }	0.024	×
再選＋目指す	0.026	×	計画＋進める	0.024	○
一般＋公開する	0.026	○	作品＋並ぶ	0.024	○
高い＋みる	0.026	×	市議＋告示する	0.024	×
改革＋進める	0.026	○	対象＋実施する	0.024	×
時間＋かける	0.026	○	地裁＋認める	0.024	×
大きい＋変わる	0.026	○	人＋知る	0.024	○
逮捕＋起訴する	0.026	×	計画＋策定する	0.023	×
まとめ＋わかる	0.026	×	事故＋調べる	0.023	×
意気込み＋語る	0.026	○	センター＋開く	0.023	×

表 14: ダイス係数 4-Xgram,MID50

共起	DC	正誤	共起	DC	正誤
酒気＋帯びる	0.755	○	現行＋逮捕する	0.193	×
耳＋傾ける	0.438	○	全力＋尽くす	0.193	○
骨＋折る	0.347	○	うまい＋いく	0.191	○
球＋振る	0.320	○	余罪＋追及する	0.189	○
熱戦＋繰り広げる	0.317	○	重傷＋折る	0.188	○
書類＋送検する	0.293	○	調べ＋よる	0.182	○
汗＋流す	0.269	○	罪＋問う	0.180	○
体調＋崩す	0.251	○	傷＋探る	0.180	×
軽傷＋負う	0.236	○	カギ＋握る	0.178	○
記者＋会見する	0.234	○	正式＋決定する	0.177	○
詐欺＋振り込める	0.234	×	身＋つける	0.174	○
酒＋飲む	0.225	○	役割＋果たす	0.171	○
アルコール＋検出する	0.223	○	口座＋振り込む	0.171	○
握手＋交わす	0.220	○	支払い＋命じる	0.170	○
ガラス＋割る	0.216	○	ウイルス＋検出する	0.166	○
１１０番＋通報する	0.215	○	協定＋結ぶ	0.165	○
原因＋調べる	0.213	○	定年＋退職する	0.163	○
重傷＋負う	0.209	○	首＋絞める	0.160	○
震度＋観測する	0.209	○	基金＋取り崩す	0.158	×
足＋運ぶ	0.207	○	出馬＋表明する	0.156	○
力＋入れる	0.203	○	胸＋張る	0.154	○
けが＋負う	0.202	○	理解＋深める	0.151	○
疑い＋逮捕する	0.202	×	回＋振る	0.147	×
工夫＋凝らす	0.200	○	立候補＋表明する	0.147	○
雨＋降る	0.195	○	笑顔＋見せる	0.144	○

表 13: ダイス係数 2-Xgram,TOP50

共起	TS	正誤	共起	TS	正誤
調べ＋よる	74.1	○	事実＋認める	28.0	○
疑い＋逮捕する	48.5	×	耳＋傾ける	27.7	○
記者＋会見する	47.9	○	事情＋聞く	27.4	○
原因＋調べる	43.7	○	疑い＋盗む	27.3	×
現行＋逮捕する	42.4	×	賠償＋求める	27.2	○
力＋入れる	41.6	○	抱負＋語る	27.1	○
容疑＋逮捕する	40.6	×	注意＋呼びかける	26.9	○
考え＋示す	37.2	○	契約＋結ぶ	26.8	○
人＋いる	36.2	○	訴訟＋求める	26.3	○
声＋かける	34.0	○	遺体＋見つかる	26.2	○
強い＋打つ	32.7	○	絵＋描く	26.1	○
酒気＋帯びる	32.6	○	方針＋固める	25.9	○
書類＋送検する	32.0	×	準備＋進める	25.8	○
笑顔＋見せる	31.8	○	酒＋飲む	25.6	○
身＋つける	31.1	○	役割＋果たす	25.5	○
足＋運ぶ	30.8	○	時間＋かかる	25.4	○
罪＋問う	30.0	○	花＋咲かせる	25.2	○
軽傷＋負う	30.0	○	人気＋集める	25.1	○
話＋聞く	29.6	○	病院＋運ぶ	25.1	○
影響＋与える	29.5	○	球＋振る	25.0	○
立候補＋表明する	29.3	×	出馬＋表明する	24.9	×
声＋上がる	29.2	○	声＋聞く	24.8	○
けが＋負う	29.0	○	影響＋出る	24.8	○
詐欺＋振り込める	28.8	×	力＋込める	24.7	○
重傷＋負う	28.2	○	熱戦＋繰り広げる	24.7	○

表 15: T スコア 2-Xgram,TOP50

ケーションとしての特徴的な表現をもたない自明な意味をもつ共起が多いことに起因する．相互情報量，ダイス係数，T スコアは全体的に精度が高く，頻度のみでは測れない，特徴的な共起を抽出できる指標であると考えられる．相互情報量では低頻度な語の影響を受けると予想されたが，頻出共起語の抽出により極端に低い頻度の語がスクリーニングされたため結果が良好になったと考えられる．対数尤度比については他の 3 種と比べて精度が低く，具体的な共起の傾向は共起頻度による抽出と似た結果であった．

4.3 連鎖共起によるコロケーション抽出精度

連鎖共起による抽出精度について，最良で 2-Xgram の 78.0%であることを示した．この連鎖共起スコアは相関ルールにおける確信度の算出と同義であるが，共起がコロケーションであると確信できるスコアはどのあたりであるかを評価する．連鎖共起のスコアに上限のしきい値をもうけ，そのしきい値未満での高スコア 50 件の正誤を判定し，抽出精度を確認する．前項の結果を踏まえ，4-Xgram までの抽出結果に対し評価を行う．この結果を図 3 に示す．

最も精度の高い 2-Xgram の TOP50 の具体例を表 19 に，精度の低い 4-Xgram のスコアしきい値上限 0.050 での具体例を表 20 表に示す．連鎖共起スコアが 0.40 未満になると常に下降する傾向を示す．これより 0.40 以上をもつ共起は結びつきが強く，コロケーションであろう特徴的な共起であることが分かる．

5 結論

本研究ではデータマイニング手法による日本語コロケーション抽出と統計指標を用いた精度評価を行った．n-Xgram にて，コロケーションとされる共起の距離は

共起	TS	正誤	共起	頻度 TS	正誤
LOC＋{ 大会, 開く }	2.54	×	相談＋受ける	0.019	○
事業＋進める	2.53	○	人＋楽しむ	-0.229	○
臨時＋開く	2.42	×	報告＋受ける	-0.505	○
人＋思う	2.33	×	問題＋考える	-0.605	○
命令＋受ける	2.27	○	試合＋行う	-0.766	○
調査＋始める	2.25	○	LOC ＋ ORG＋行う	-0.890	×
問題＋めぐる	2.05	○	大会＋目指す	-1.44	○
イベント＋開く	2.05	○	地方＋よる	-1.74	×
LOC ＋ ORG＋よる	1.99	×	人＋来る	-1.86	○
笑顔＋話す	1.85	○	記者＋開く	-1.89	×
教委＋よる	1.69	×	問題＋巡る	-1.94	○
治療＋受ける	1.61	○	問題＋解決する	-2.12	○
PER＋{ 立候補, 表明する }	1.53	×	LOC＋{PER, 訪れる }	-2.16	○
作品＋見る	1.44	×	PER ＋ 強い＋打つ	-2.20	×
市長＋話す	1.40	×	講座＋開く	-2.41	○
容疑＋供述する	1.33	×	LOC＋{ 訴訟, 求める }	-2.64	×
活動＋行う	1.17	○	男性＋話す	-2.69	○
ホール＋開く	1.05	×	現場＋よる	-2.79	×
同社＋よる	0.870	×	利用＋できる	-2.79	×
予定＋開く	0.710	×	監督＋話す	-2.86	○
PER＋{LOC, 務める }	0.650	×	社長＋話す	-2.92	○
問題＋取り組む	0.640	○	市＋受ける	-3.02	×
午前＋よる	0.580	×	問題＋抱える	-3.13	○
教授＋話す	0.220	×	関係＋よる	-3.49	×
市民＋開く	0.100	×	要請＋受ける	-3.49	○

表 16: T スコア 4-Xgram,MID50

共起	LLR	正誤	共起	LLR	正誤
PER＋よる	17.4	×	どう＋する	14.6	×
PER＋話す	17.2	○	LOC＋受ける	14.5	×
明らか＋する	17.1	○	記者＋会見する	14.5	○
調べ＋よる	16.8	○	可能＋なる	14.4	×
可能＋ある	16.8	×	中心＋する	14.4	○
LOC＋ある	16.5	×	見通し＋なる	14.3	×
LOC＋する	16.5	×	原因＋調べる	14.2	○
PER＋する	16.2	×	問題＋ある	14.2	○
LOC＋なる	16.1	×	現行＋逮捕する	14.2	×
必要＋ある	16.1	○	PER＋務める	14.2	○
LOC＋開く	16.0	×	大切＋する	14.2	○
LOC＋行う	15.7	×	ORG＋なる	14.2	×
LOC＋よる	15.5	×	安定＋する	14.1	○
対象＋する	15.5	○	LOC＋訪れる	14.1	○
ORG＋よる	15.5	×	容疑＋認める	14.1	○
容疑＋逮捕する	15.4	○	考え＋示す	14.1	○
ORG＋する	15.2	×	選挙＋なる	14.1	×
PER＋なる	15.1	×	力＋入れる	14.1	○
人＋いる	15.1	○	ORG＋ある	14.1	×
処分＋する	15.0	○	仕事＋する	14.1	○
明らか＋なる	15.0	○	判決＋よる	14.0	×
疑い＋逮捕する	14.7	○	PER＋語る	14.0	○
恐れ＋ある	14.7	○	必要＋する	14.0	×
PER＋ある	14.7	×	人＋する	14.0	×
対象＋なる	14.7	○	PER＋受ける	14.0	×

表 17: 対数尤度比 2-Xgram,TOP50

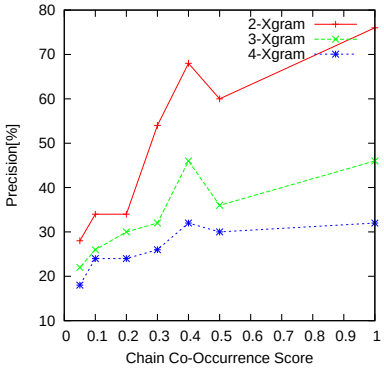

図 3: 連鎖共起スコアに対する精度

共起	LLR	正誤	共起	LLR	正誤
地方 ＋ LOC＋よる	11.8	×	投票＋する	11.8	○
LOC＋楽しむ	11.8	×	動き＋ある	11.8	○
LOC＋守る	11.8	×	大会＋なる	11.8	×
4月＋よる	11.8	×	LOC ＋ PER＋行う	11.8	×
結果＋発表する	11.8	○	男性＋話す	11.8	○
PER＋始まる	11.8	×	選挙＋ある	11.7	×
LOC＋設置する	11.8	×	LOC＋示す	11.7	×
昨年＋する	11.8	×	判決＋{ 被告,PER, よる }	11.7	×
PER＋立候補する	11.8	×	チーム＋する	11.7	×
会社＋ある	11.8	×	国＋する	11.7	×
病院＋する	11.8	×	報告＋受ける	11.7	○
交流＋深める	11.8	○	PER＋亡くなる	11.7	×
自分＋ある	11.8	×	市長＋なる	11.7	×
地裁＋求める	11.8	×	事業＋ある	11.7	×
余儀ない＋する	11.8	○	監督＋話す	11.7	○
制度＋ある	11.8	×	午前＋行う	11.7	×
ほか＋なる	11.8	×	傷害＋逮捕する	11.7	×
犠牲＋なる	11.8	○	LOC＋開幕する	11.7	×
舞台＋する	11.8	×	PER＋活躍する	11.7	×
PER＋{LOC, 訪れる }	11.8	×	準備＋する	11.7	×
障害＋ある	11.8	×	力＋注ぐ	11.7	○
事情＋ある	11.8	×	社長＋話す	11.7	○
観光＋訪れる	11.8	○	いい＋思う	11.7	○
試合＋なる	11.8	×	PER＋出場する	11.7	○
確認＋急ぐ	11.8	○	PER＋あいさつする	11.7	×

表 18: 対数尤度比 4-Xgram,MID50

共起	頻度 []	正誤	共起	頻度 []	正誤
する ｜ 余儀ない	0.997	○	工夫 ｜ 凝らす	0.605	○
帯びる ｜ 酒気	0.997	○	負う ｜ 軽傷	0.593	○
向かう ｜ 快方	0.994	○	酒 ｜ 酔う	0.590	○
詐欺 ｜ 振り込める	0.956	×	よる ｜ 気象台	0.579	×
目 ｜ 細める	0.933	○	病院 ｜ 搬送する	0.578	○
花 ｜ 咲かせる	0.896	○	する ｜ はっきり	0.571	○
PER ｜ 承る	0.896	×	懲役 ｜ 求刑する	0.570	×
なる ｜ お世話	0.895	○	負う ｜ やけど	0.564	○
記者 ｜ 会見する	0.889	○	語る ｜ 抱負	0.551	○
する ｜ ホッ	0.830	×	誓う ｜ 健闘	0.550	○
する ｜ ほっと	0.815	○	繰り広げる ｜ 熱戦	0.545	○
なる ｜ 一丸	0.806	○	容疑 ｜ 否認する	0.545	○
耳 ｜ 傾ける	0.804	○	できる ｜ 納得	0.538	○
なる ｜ 励み	0.768	○	骨 ｜ 折る	0.527	○
酒気 ｜ 帯びる	0.755	○	気 ｜ 引き締める	0.516	○
ある ｜ 恐れ	0.729	○	なる ｜ 通行止め	0.508	×
基金 ｜ 取り崩す	0.727	×	流す ｜ 汗	0.494	○
送る ｜ エール	0.711	○	問う ｜ 罪	0.493	○
する ｜ 明らか	0.699	○	負う ｜ 重傷	0.482	○
首 ｜ 絞める	0.698	○	よる ｜ 調べ	0.471	×
なる ｜ 浮き彫り	0.675	○	逮捕する ｜ 現行	0.467	×
影響 ｜ 及ぼす	0.665	○	かける ｜ 迷惑	0.466	○
上げる ｜ 名乗り	0.640	○	被害 ｜ 遭う	0.463	○
よる ｜ 訴状	0.634	×	振る ｜ 回	0.455	×
書類 ｜ 送検する	0.622	○	する ｜ 題材	0.452	○

表 19: 連鎖共起 2-Xgram, 上限しきい値 1.0

自立語 4 語以内にみられることを示した．統計指標による抽出は相互情報量，ダイス係数，T スコアがすぐれており，その抽出精度は最大で 88.0%であることを示した．連鎖共起における抽出精度はそのスコアが 0.40 以上のときに好ましく，コロケーションであろう特徴的な共起であることを示した．

共起	スコア	正誤	共起	スコア	正誤
PER ｜ 見る	0.050	×	影響 ｜ 出る	0.049	○
する ｜ 一つ	0.050	×	なる ｜ 初めて	0.049	×
する ｜ 効果	0.050	×	PER ｜ 伴う	0.049	×
LOC ｜ 出す	0.050	×	LOC＋開く ｜ ホテル	0.048	×
ORG ｜ {LOC, する }	0.050	×	出火 ｜ 調べる	0.048	×
する ｜ 意見	0.050	×	理由 ｜ 説明する	0.048	○
求める ｜ 地裁	0.050	×	説明する ｜ 理由	0.048	×
受ける ｜ 判決	0.050	×	巡る ｜ 工事	0.048	×
ある ｜ 気持ち	0.049	×	{ 調べ,PER} ｜ よる	0.048	×
注目 ｜ 集まる	0.049	○	よる ｜ 現場	0.048	×
声 ｜ 相次ぐ	0.049	×	する ｜ 問題	0.048	×
事件 ｜ 調べる	0.049	○	調べ ｜ わかる	0.048	×
する ｜ 強い	0.049	×	なる ｜ 場合	0.048	×
視野 ｜ 入れる	0.049	○	全焼する ｜ 住宅	0.048	×
行う ｜ 練習	0.049	×	ある ｜ 状態	0.048	×
運ぶ ｜ 病院	0.049	×	LOC ｜ 話す	0.048	×
する ｜ 相手	0.049	×	得る ｜ 支持	0.048	○
込める ｜ 思い	0.049	○	示す ｜ 見通し	0.047	×
迷惑 ｜ かける	0.049	×	帯びる ｜ 運転	0.047	×
費用 ｜ かかる	0.049	×	LOC ｜ 調べる	0.047	×
ない ｜ 変わる	0.049	×	更新する ｜ 最高	0.047	×
ORG ｜ 目指す	0.049	×	{ 疑い, 容疑 } ｜ 逮捕する	0.047	×
講師 ｜ 務める	0.049	×	ORG ｜ PER＋する	0.047	×
励む ｜ 練習	0.049	○	PER ｜ 学ぶ	0.047	×
受ける ｜ 被害	0.049	○	ある ｜ 障害	0.047	×

表 20: 連鎖共起 4-Xgram, 上限しきい値 0.05

参考文献

[1] Stubbs,M: Words and Phrases - Corpus Studies of Lexical Semantics,Blackwell Publishers,2001

[2] 園田匠 , 三浦孝夫: 日本語コロケーションのためのデータマイニング , 第 4 回データ工学と情報マネジメントに関するフォーラム,2012

[3] Justeson, J., Katz, S.: Technical terminology: some linguistic properties and an algorithm for identi cation in text. Natural Language Engineering,1995

[4] Han, J. and Kamber, M.: Data Mining, Morgan Fauffman, 2006

[5] 石川慎一郎 : 英語コーパスと言語教育―データとしてのテクスト,2008, 大修館書店

[6] 姫野昌子: 日本語表現活用辞典, 2004, 研究社