7Q-03

入れ子構造に着目した動詞コロケーションの段階的獲得

町田　翔*　　河野　一志*　　村松　拓実*　　延澤　志保*
*東京都市大学知識工学部

1　はじめに

コロケーションとは習慣的に用いられる自然な語の組み合わせである [1]．コロケーションを網羅的に収集することによって，文章の推敲や校正を自動的に行うための資源としての用途が考えられる [2]．ある動詞に対して共起する語との自然な語の組み合わせを動詞コロケーションという．例えば｛聞く｝は数多くの語と共起するが，その中でも｛澄まし，て，聞く｝は自然な語の組み合わせである．また,｛澄まし, て, 聞く｝は｛耳, を｝が先行する場合が多いように，いくつかの語の連なりに対して周辺語句が限定される場合，その語の組み合わせはひとまとまりの動詞句として考えられる.｛耳, を, 澄まし, て, 聞く｝は｛じっと，耳，を，澄まし，て，聞く｝のように，ある動詞句に対して共起する語との自然な組み合わせも存在する．よって，動詞コロケーションは語と語の共起関係だけでなく，語と句の共起関係も収集すべきである．そこで本稿では，動詞コロケーションを段階的に獲得する手法を提案する．

2　関連研究

園田らは Stubbs の英語コロケーションの分類基準を日本語に適用させたコロケーション抽出を行った [1]．日本語と英語の違いとして，日本語は意味が語順ではなく格助詞に依存して決定することから，自立語間の共起確率を用いて日本語コロケーション抽出をしている．しかし，2 つの自立語間の共起確率しか捉えていないため動詞句｛耳, を, 傾け, て, 聞く｝を抽出することができない．

新納らは 2 単語間の結合度を測る自己相互情報量 (Pointwise Mutual Information, PMI) を拡張し,｛かたず, を, 飲む｝などの｛名詞，助詞，動詞｝で構成されている動詞句の抽出を行った [3]．PMI 値の高い組み合わせは，語と語が同時に出現しやすいとができる [5]．しかし，4 単語以上で構成されている動詞句も存在するため，動詞句を構成する語数に囚われない動詞句の獲得手法が必要である．

Stepwise Acquisition of Verbal Collocations Considering Nest Structure
Sho Machida*, Kazushi Kouno*, Takumi Muramatsu*, and Shiho H.Nobesawa*.
* Faculty of Knowledge Engineering, Tokyo City University

3　動詞コロケーションの段階的獲得

3.1　提案手法の概要

動詞コロケーションを抽出するために，語と語の結合力が強い組み合わせを動詞句として獲得する．しかし，動詞句はいくつかの語の組み合わせであり，動詞句を構成する語数が明確でない．そこで，結合力の強い 2，3 語の組み合わせに対して高い確率で先行後続する語は，その組み合わせの一部である可能性が高いことから，2，3 語の動詞句から段階的に構成語を追加し動詞句を獲得する．獲得した動詞句を対して先行する語の品詞情報を用いてコロケーション抽出する手法を提案する．

3.2　動詞句の獲得

語と語の結合力を測る尺度として，新納らは PMI に基づく手法を提案している [3]．2 つの要素 x, y に対して，PMI は式 (1) で定義される [5].

$$PMI(y,z) = \log_2 \frac{P(y,z)}{P(y)P(z)} \tag{1}$$

$P(y,z)$ は要素 y と z が同時に出現する確率，$P(y)$, $P(z)$ はそれぞれ要素 y，z が出現する確率である．PMI 値が正のとき，y と z が同時に出現しやすいことを示し，負のときは同時に出現しにくいこと示す．しかし，式 (1) の 2 単語間の結合力だけでは 3 単語で構成されている動詞句の結合力を測れない．そのため (x,y,z) で出現している場合，yz をひとまとまりの語であると仮定し，先行する語 x と yz の結合力を測る式を式 (2) で定義する．

$$PMI(x,yz) = \log_2 \frac{P(x,yz)}{P(x)P(yz)} \tag{2}$$

x は動詞句 (y,z) に先行する語であり，yz は動詞句である．(y,z) に対して先行する語 x との結合力を測ることにより,｛念, を, 覚える｝などの動詞句の結合力が強くなると考えられる．この式 (2) を用いて trigram の PMI 値を算出し，式 (1) で算出した bigram の PMI 値と比較する．PMI 値が 1 以上の組み合わせを抽出し，比較した結果の一部を表 1 に示す.

表 1 から，bigram の方が trigram より PMI 値が高い場合，bigram の動詞句である可能性が高く，trigram の方が bigram より PMI 値が高い場合は，trigram 以

表 1: 動詞 bigram と動詞 trigram の比較

動詞句 yz	PMI 値		動詞句 xyz	PMI 値
習う覚える	8.874	>	を習う覚える	3.434
はっきり覚える	6.755	>	ははっきり覚える	2.259
を覚える	4.082	>	だけを覚える	2.142
			悪寒を覚える	10.545
を覚える	4.082	<	快感を覚える	10.472
			念を覚える	7.372

上で構成されている動詞句である可能性が高いと考えられる．

獲得した trigram を対象に，どこからどこまでが動詞句かを段階的に調査する．動詞句 (s) の構成要素を $w_i...w_j$ と仮定し，s の構成要素数を $length(s)$ とする．s に対して w_{i-1} が一定の確率で共起する場合，$length(s)+1$ 以上で構成されている動詞句であると考えられ，s の構成要素を $w_{i-1}...w_j$ とし，s に対して w_{i-2} を調査する．s に対して w_{i-1} が一定の確率未満の場合，$length(s)$ で構成されている動詞句として扱う．この処理をすべての動詞句を獲得するまで続ける．後続する語 w_{j+1} に対しても同様の処理をし，動詞句 s を獲得する．この処理を用いて獲得した動詞句の一部を表 2 に示す．

表 2: 動詞句 (trigram) として抽出した語句の一部

胸が騒ぐ	肩を並べて歩く	犬馬の労をつくす
腹を満たす	飛ぶ鳥を落とす	合槌を打つ
ひとたまりもない	の目を盗む	耳を澄まして聞く
群を抜く	神経衰弱にかかる	影も形も見せるない
今も覚えるている	の念を覚える	矢も盾もたまるぬ

3.3 動詞コロケーションの抽出

動詞コロケーションを抽出する際，動詞，動詞句に先行する語の品詞を定める．動詞，動詞句に先行する語の品詞を調査した結果,{名詞, 助詞}{副詞}{名詞, 助詞, 助詞}{名詞}{接頭詞}{動詞}{形容詞} の品詞の組み合わせがコロケーションであると判断した.{名詞}{動詞}{形容詞}{副詞} は独立して意味を持ち，動詞に対し先行しやすい語であるため獲得すべきであり,{接頭詞} は自立語に前接して全体で一語を形成して文法的特徴を加える役割を持っているため獲得すべきである．動詞コロケーションとして，獲得した動詞句と動詞句に先行する定めた品詞との組み合わせを抽出した．

4 実験結果と考察

提案手法を用いて青空文庫コーパス [4] の 2,526,860 文を使用し，動詞コロケーションの獲得を行った．動詞句の獲得の際，動詞句 $(w_i...w_j)$ に対して先行後続する語の出現確率の閾値は 90%とし，bigram は 1,681 語，trigram は 22,634 語，4gram 以上は 660 語，すべて合わせて 24,975 語の動詞句を獲得した．これらの動詞句に対し，品詞情報を用いてコロケーションを抽出した．抽出したコロケーションの一部を表 3 に示す．

表 3: 獲得したコロケーションの一部

覚える	よく覚える	不安を覚える	の念を覚える	の情を覚える
ハッキリ	いろいろ	まだ	尊敬	憤懣
情を	不思議に	将来に対する	後悔	嫌悪
念を	顔を	急に	同情	反撥
不安を	最も	仕事に	感謝	安心
礼儀も	時は	一種の	憎悪	恋慕

評価方法として，青空文庫コーパス [4] から無作為に選択した 1000 文に対して獲得した動詞コロケーションを抽出した．正解データはコロケーション辞典 [6] に記載されている語の組み合わせとした．抽出された動詞コロケーションを正解データを用いて評価した結果が表 4 である．

表 4: 提案手法の抽出精度

ngram	正解率
bigram	58%
trigram	64%
ngram(n>3)	38%

trigram の正解率が 64%と有効性が確認できた．言い回しとして正解であるコロケーションも辞書に記載されていないために不正解となった表現があったためと考えられる.「犬馬の労をつくす」は「犬馬の労をとる」と記載されており，出現していない語の組み合わせを獲得できないため不正解となった．

5 まとめ

本研究では，入れ子構造に着目した動詞コロケーションの段階的獲得手法を用いることで，確率的に動詞句を獲得でき，動詞コロケーションを抽出することができた．段階的に獲得することにより，動詞句の構成語数に囚われない抽出の有効性を示すことができた．

参考文献

[1] 園田匠, 島田諭, 三浦孝夫, “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,” 第 11 回情報科学技術フォーラム, E-021, pp.201–206, 2012.

[2] 田野村忠温, “コーパスからのコロケーション情報抽出ー分析手法の検討とコロケーション辞典項目の試作ー,” 阪大日本語研究会, pp.21–41, 2009.

[3] 新納浩幸, 井佐原均, “片方共起による述語型定型表現の自動抽出,” 言語処理学会, 自然言語処理, Vol.2, No.3, pp.73–86, 1995.

[4] 青空文庫, http://www.aozora.gr.jp/

[5] 奥村学, “言語処理のための機械学習入門,” コロナ社, 2015.

[6] 小内一, “てにをは辞典,” 三省堂, 2010.