大規模コーパスからの可変長 n 項コロケーション自動抽出

Automatic extraction of variable-length n-gram collocations from a large corpus

内村英彦† 森山秀‡ 二村良彦‡‡
早稲田大学大学院理工学研究科情報科学専攻† 早稲田大学理工学部複合領域‡ 早稲田大学理工学部情報学科‡‡

1 はじめに

第 2 言語学習者にとってコロケーションの理解・習得は困難とされており、コロケーション自動抽出法の研究は機械翻訳だけではなく第 2 言語学習法への応用面でも期待されている。本論では自然言語に見られる共起現象の再帰性に着目し、単語列の長さに制限を設けない可変長 n 項コロケーションをモデル化する。一般的には、コーパスから n 項モデルに基づく情報抽出を行う時、空間計算量が最大の問題となる [2, pp.192-195]。本論で示すモデルに基づき、品詞情報つき大規模コーパスからそのサイズの空間計算量で可変長 n 項コロケーションを自動抽出する方法を提案する。

2 可変長 n 項コロケーションモデル

$W(\varepsilon \in W)$ をコーパス中に出現する語彙集合とする。また、w_i と表す時、語 w_i は語 w_0 から相対位置 i にあることを意味するものとする[1]。$C(w_i|w_0)$ を語 w_0 と相対位置 i で語 w_i が共起する 2 項コロケーション、$C(w_{\pm i,\ldots\pm 1}|w_0)$ を語 w_0 と語の列 $w_{\pm i,\ldots\pm 1}$ が w_0 の前方または後方に位置して共起する n 項コロケーション、$C(w_{-i,\ldots-1}, w_{1,\ldots j}|w_0)$ を語 w_0 と語の列 $w_{-i,\ldots-1}, w_{1,\ldots j}$ が w_0 の前後方に位置して共起する n 項コロケーションを表すものとする[2]。以下は辞書 [1] に見られるコロケーションと $C(\ldots)$ による表記法の例である[3]。

unexpected news	$C(unexpected\|news)$
in the line of duty	$C(in\ the\ line\ of\|duty)$
come into conflict with	$C(come\ into,\ with\|conflict)$

文脈自由文法を基礎とする可変長 n 項コロケーションモデルを構築する。開始記号を、w_0 を中心としたコロケーションを意味する $C(w_0)$ とし、終端記号を W、非終端記号を $\{C(w)|w \in W\}$ とする。また生成規則を、

$$C(w_i) \to C(w_i)w_j|C(w_i)C(w_j)|w_i \ (i < j)$$

$$\to w_jC(w_i)|C(w_j)C(w_i)|w_i \ (i > j)$$

と定義する。コーパスから抽出された 2 項コロケーションから生成規則を導出 ($\Rightarrow$) する[4]。例として $i > 0$ の時

$$C(w_i|w_0) \Rightarrow C(w_0) \to C(w_0)w_i$$

と、語とコロケーションの共起関係を表す生成規則が得られる。また w_0 に対して相対位置 i に出現する w_i を中心とするコロケーションが存在する場合、

$$C(w_i|w_0) \Rightarrow C(w_0) \to C(w_0)C(w_i)$$

と、コロケーション間の共起関係を表す生成規則が得られる。

ここで生成規則の適用 ($\to$) 毎に語同士の相対位置（添字）に基づいて整列 ($\overset{sort}{\to}$) を行うよう文脈自由文法を拡張する。以下は、この拡張された文法による可変長 n 項コロケーションの導出過程の例である。

$$\begin{aligned} C(w_0) &\to C(w_0)w_i \overset{sort}{\to} w_iC(w_0) \ (i < 0) \\ &\to w_iC(w_0)C(w_j) \\ &\to w_iC(w_0)w_kC(w_j) \overset{sort}{\to} w_iC(w_0)C(w_j)w_k \\ &\quad (i < 0 < j < k) \\ &\to \ldots \to \ w_{-l,\ldots-1}w_0, w_{1,\ldots m} \\ &= C(w_{-l,\ldots-1}, w_{1,\ldots m}|w_0) \end{aligned}$$

3 可変長 n 項コロケーション抽出

頻繁に、かつ偶然以上の確率で共起する語の組み合わせをコロケーションと定義する。$f(x)$ を x の出現する頻度の関数とし、T_f を出現頻度の閾値とすれば、任意のコロケーション C は $f(C) \geq T_f$ を満たさなければならない。また、$d(x)$ を共起関係の強さの判定関数とし、T_d を共起関係の強さの閾値とすれば、任意のコロケーション C は $d(C) \geq T_d$ を満たさなければならない。

本研究では可変長 n 項コロケーションを 3 段階の処理を経て抽出する。

3.1 2 項コロケーション抽出

コーパスから位置関係を考慮した語の 2 項関係を抽出する方法は [2, 3, 5] 等に詳細に記されている。本研究では出現頻度の条件を満たしかつ、2 項関係における共起関係の強さの判定関数 $d(C(w_i|w_0))$ を

$$\log \frac{p(w_i|w_0)}{p(w_i)} - \delta\big(\log \frac{p(t_i|t_0)}{p(t_i)}\big)$$

† Hidehiko Uchimura
† Department of Information and Computer Science, Science and Engineering, Waseda University Graduate School
‡ Hiizu Moriyama
‡ Multidisciplinary Studies, Science and Engineering, Waseda University
‡‡ Yoshihiko Futamura
‡‡ Department of Information and Computer Science, Science and Engineering, Waseda University

[1] 添字に対して特に条件を記述しない場合、添字 $i(\neq 0) \in Z$ とする。
[2] 語の列の添字を記述する場合のみ、コロケーションの中心の語との相対位置関係を強調するため、添字 $i(\neq 0) \in N$ とする。
[3] 語の品詞情報は省略。
[4] 「$c \Rightarrow r$」は 2 項コロケーション c がコーパス中に存在する時、r は生成規則であることを意味する。

とした時、条件 $d(C(w_i|w_0)) \geq T_f$ を満たすものの集合をコーパスから抽出する。なお、これ以降 $p(x)$ は事象 x がコーパス中で起こる確率とし、t_i を語 w_i に対応する品詞とし、

$$\delta(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

とする。$\delta(x)$ は文法のみに依存して共起する語の組み合わせを除くことを目的とした関数である[5]。

3.2 可変長 n 項コロケーション候補抽出

語 w_0 を中心としたコロケーション集合を抽出する時、w_0 を含む文全体からなるコーパスの部分集合 $S(w_0)$ を処理の対象領域と考えてよい。$len(x)$ を語の列 x の長さの関数、$C(w_0, s)$ を $s(\in S(w_0))$ 中で可変長 n 項コロケーションモデルにより認識される語の列の集合とする。コーパスから抽出される w_0 を中心とした可変長 n 項コロケーション候補集合 $CC(w_0)$ は以下のとおり定義される。

$$CC(w_0) = \{\arg \max_{c \in C(w_0,s)} len(c) | s \in S(w_0)\}$$

この段階では $CC(w_0)$ の要素はコロケーション判定基準を満たしているかどうかわからない。しかし、w_0 を中心とするコロケーション c がコーパス中に存在し、コロケーションを語の列と考える時、c を部分列とする $CC(w_0)$ の要素が存在する[6]。よって、この集合を可変長 n 項コロケーション候補集合とする。

3.3 可変長 n 項コロケーション抽出

2 つの木を有する森を $CC(w_0)$ から構築する。木のパスは語の列に対応している。深さ i のノードは語 $w_{\pm i}$ に対応し、そのノードまでのパスに対応するコロケーション候補の出現頻度を持つ。一方の木のパスは w_0 を中心として前方に、もう一方の木のパスは後方に共起する語の列を表す。

この森に対してコロケーションの条件を適用し、木のノードの合併および枝刈を行う。木に対しては共起関係の強さの判定関数 $d(C(w_{\pm i,\ldots\pm 1}|w_0))$ を

$$\log \frac{p(w_{\pm i,\ldots\pm 1}|w_0)}{p(w_{\pm(i-1),\ldots\pm 1}|w_0)p(w_{\pm i})} - \delta\Big(\log \frac{\min_{0<i'<i} p(t_{\pm i'}, t_{\pm i}|t_0)}{\min_{0<i'\leq i-1} p(t_{\pm i'}|t_0)p(t_{\pm i})}\Big)$$

とする。森つまり、w_0 を中心として前後方に共起するコロケーションに対して共起関係の強さの判定関数 $d(C(w_{-i,\ldots-1}, w_{1,\ldots j}|w_0))$ は以下のとおりである。

$$\log \frac{p(w_{-i,\ldots-1}, w_{1,\ldots j}|w_0)}{p(w_{-i,\ldots-1}|w_0)p(w_{1,\ldots j}|w_0)} - \delta\Big(\log \frac{\min_{-i\leq i'<0<j'\leq j} p(t_{i'}, t_{j'}|t_0)}{\min_{-i\leq i'<0} p(t_{i'}|t_0) \min_{0<j'\leq j} p(t_{j'}|t_0)}\Big)$$

コロケーション条件の適用により縮小された森を $CC(w_0)$ のそれぞれの要素の語の列に従って探索する。2 つの木それぞれの探索によって得られたパスを連結することにより 1 つの w_0 を中心とする可変長 n 項コロケーションが得られる。

4 実験

前節で提案した共起関係の強さの判定関数 d を用いた可変長 n 項コロケーション抽出法の妥当性を検証するために、品詞情報つき大規模コーパスとして代表的な存在である BNC[7] から可変長 n 項コロケーションを抽出した。本研究の目的の一つは英語学習者に有用なコロケーション情報の自動抽出である。よって、コロケーション辞書として定評があり過去の研究 [5] でも使用されている [1] からサンプルをとり、評価実験を行った。その結果、サンプルに代表されるコロケーションは $72.3 \pm 3.0\%$ の比率で抽出されると推定された。[8]。

5 まとめ

本研究ではコロケーションを語と語の共起関係だけではなく、語とコロケーション間、さらにコロケーション間と再帰的に考えることにより共起する語の列の長さを 2 項から任意の長さに拡張し、柔軟なコロケーション自動抽出法を提案した。この方法ではコロケーションの中心となる語に焦点を当て抽出処理を行う。これはコーパスからのコロケーション検索に応用可能な方法であり、より英語学習者にとって有用なソフトウェア開発を可能にすることが期待できる。

参考文献

[1] Benson, M. et al. (1993), The BBI combinatory dictionary of English, Amsterdam: Benjamins.

[2] Manning, C. D. and H. Schutze. (2002), Foundation of Statistical Natural Language Processing 5th ed., Massachusetts: The MIT Press.

[3] Oakes, M. P. (1998), Statistics for Corpus Linguistics, Edinburgh: Edinburgh University Press.

[4] Sinclair, J. (1991), Corpus Concordance Collocation, Oxford: OUP.

[5] Smadja, F. (1993), ‘ Retrieving Collocations from Text: Xtract ’ , Computational Linguistics, 19(1), pp. 143-177.

[5] 品詞列「冠詞」+「名詞」には語に依存しない強い共起関係がある。

[6] $\varepsilon(\in W)$ は同位置の任意の語と適合するものとする。

[7] http://www.hcu.ox.ac.uk/BNC/

[8] [5] にあるように、[1] による評価実験では 80% 前後の抽出率がひとつの目安となる。本実験でそれ以下の抽出率になった理由の一つは、文法のみに依存して共起する語の組み合わせをコロケーションと考えないことがあげられる。なお、サンプルサイズは 800 以上、推定の信頼度は 95%、閾値 $T_f = 8$, $T_d = 1.0$ である。