Web文書中の語の共起性を用いたコロケーション翻訳支援システムの実装

柴田, 雅博
九州大学大学院システム情報科学府知能システム学専攻 : 博士後期課程

冨浦, 洋一
九州大学大学院システム情報科学研究院知能システム学部門

田中, 省作
情報基盤センター外国語情報メディア研究部門

https://doi.org/10.15017/1516059

出版情報：九州大学大学院システム情報科学紀要. 10 (1), pp.39-44, 2005-03-25. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学紀要
第10巻 第1号 平成17年3月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.10, No.1, March 2005

Web文書中の語の共起性を用いたコロケーション翻訳支援システムの実装

柴田 雅博* ・ 冨浦 洋一** ・ 田中 省作***

Implementation of Writing Support System for Translating Collocations Based on the Word Co-occurrence on the Web Documents

Masahiro SHIBATA, Yoichi TOMIURA and Shosaku TANAKA

(Received December 24, 2004)

Abstract: This paper explains the implementation of the system for retrieving candidates for the translation of collocations (combinations of words) from the web. For non-native speakers like Japanese, writing in English is a hard work in order not to use unnatural collocations. Sometimes they have to consult dictionaries or real documents to judge naturalness of English collocations. Besides, it is also difficult to seek out the proper candidates even using dictionaries or real documents. The proposed system is effective to seek out the proper translation of such a collocation. When a user inputs a Japanese collocation "n^J WO v^J" and gathers some nouns each of which co-occurs strongly with v^J as the same meaning as v^J in "n^J WO v^J", the system retrieves candidates for the translation of v^J from the web with their priorities. After that, the user checks validity of each candidate using the real document which the system provides and includes the English phrase consisting of a predicate verb V^E and its object n^E. Here, n^E is the translation of n^J and V^E is the candidates for the translation of v^J.

Keywords: Writing support system, Word co-occurrence, Collocation

1. はじめに

我々日本人のように英語を母語としない者にとって，英語で文章を書くのは非常に労力を要する作業である．母語以外の言語を運用するためには，文法や個々の単語の意味だけでなく，自然な語と語との組み合わせ（コロケーション†1）に十分気を配る必要がある．語彙知識の乏しい非母語話者にとって，コロケーションが自然かどうかを判断するには辞書や英語実文書に頼るほかないが，実際にコロケーションの妥当性を正確に判断するのは難しい．たとえば日本語コロケーション「ベクトル空間を張る」に対する英訳を考える．『ベクトル空間』に対する英訳が"vector space"であることは辞書を調べれば容易に分かる．それに対し，和英辞書で『張る』の訳語を調べると，"stretch", "pitch", "stick", "extend", "cover"など様々な候補が得られる．これらの候補から，辞書に記載されている例文を参考に，「ベクトル空間を張る」における『張る』の適切な訳語を求めるのであるが，記載されている少数の例から，これを判断することは一般に困難である．さらに，「ベクトル空間を張る」における『張る』に対しては，"construct", "define", "create"といった訳語の方がより妥当だと思われるが，実はこれらの単語は辞書の『張る』の項には載っておらず，このような場合，辞書を引いても妥当な訳語を得ることができない．シソーラスなどを用いて『張る』の類語まで広げて訳語を調べることで訳語候補を増やすことも考えられるが，やはり，適切なものがどれであるかを例文を頼りに判断するのは困難である．そもそも，類義語の範囲を広げすぎると「ベクトル空間を張る」の意味を保存できなくなる可能性もある．

また，英語文書から，自分の表現したい内容と類似した内容を表現している箇所を探し，その表現を参考にして辞書などを活用しながら妥当な訳語を求める場合，もしこの方法で訳語を見つけることができれば，その訳語は高い信頼性を持つ．しかし，この作業を人手で行なうのには相当な労力と時間が必要となる．

これに対し，web上の文書を実文書として扱い[1]，web上における単語の共起性に基づいて，コロケーションに対する妥当な訳語候補を半自動的に抽出する手法を提案している[2]．本稿では，提案手法を実アプリケーションとして実装し，英文書作成支援のためのコロケーション翻訳候補を半自動的に提示するシステムの構築について説明する．本システムでは，入力された日本語コロケーションに対して，提案手法にしたがって，web文書から共起データを収集し，単語の共起性に基づいて訳語候補の順

平成16年12月24日受付

* 知能システム学専攻博士後期課程
** 知能システム学部門
*** 情報基盤センター外国語情報メディア研究部門

†1 本稿では，係り受け関係 f での係り受け構造における，係る語 w と係られる語 w' の組み合わせをコロケーションと呼ぶ．なお，f としては，日本語では格助詞，英語では **subj**, **obj**, 前置詞といった表層的なものを想定している．

序付けを行ない，利用者に提示する．また，各訳語候補には，その訳語候補が実際に使われているwebページへのハイパーリンクも共に提示され，システムが提示した訳語候補が妥当かどうかをチェックするのに利用することができる．

なお，本稿では，日本語の動詞 v^J と名詞 n^J が「を」格で繋がった日本語コロケーション「n^J を v^J」に対して，その句を英語の動詞 v^E とその目的語 n^E から成るコロケーションへと翻訳する場合について議論する．これは，検索エンジンを用いて英語コロケーションを抽出する際に，パターンマッチだけでも比較的正しく係り受け関係を抜き出せることによる．

本稿の構成は以下の通りである．まず，2.章にて，提案手法の概要と基本アイデアについて説明し，3.章にて，システムの実装について述べる．

2. 提 案 手 法

2.1 共起性

単語 w が関係 f で単語 w' に係っているという係り受け構造をコロケーションと呼び，$\langle w, f, w' \rangle$ と表記する．また，コロケーション $\langle w, f, w' \rangle$ に対する w と w' との相関の強さを $\langle w, f, w' \rangle$ の共起性と呼び，$C(w, w' \mid f)$ と書く．共起性 $C(w, w' \mid f)$ は，文献 3) などで用いられている相互情報量に基づく値

$$C(w, w' \mid f) = \log \frac{P(\langle w, f, w' \rangle \mid f)}{P(\langle w, f, * \rangle \mid f) \cdot P(\langle *, f, w' \rangle \mid f)} \quad (1)$$

を想定する．ここで，$P(\langle w, f, w' \rangle \mid f)$ は係り受け関係が f であるときのコロケーション $\langle w, f, w' \rangle$ の条件付発生確率である．また，$P(\langle w, f, * \rangle \mid f)$, $P(\langle *, f, w' \rangle \mid f)$ はそれぞれ

$$P(\langle w, f, * \rangle \mid f) = \sum_{w'} P(\langle w, f, w' \rangle \mid f),$$
$$P(\langle *, f, w' \rangle \mid f) = \sum_{w} P(\langle w, f, w' \rangle \mid f)$$

である．つまり，$C(w, w' \mid f)$ は f を介したコロケーションにおける w と w' の発生の独立性からのずれを数量化したものである．

2.2 基本アイデア

提案手法は次の二つの仮定に基づいている†2．

[仮定1] 日本語名詞 n_1^J, n_2^J, $\cdots$, n_k^J に対して，コロケーション「n_i^J を v^J」($i = 1, 2, \cdots, k$) における v^J の意味が同一ならば，それらの日本語コロケーションの英訳として適切な v^J の訳語も同一である傾向にある．

[仮定2] $\langle n^J, 『を』, v^J \rangle$ の適切な英訳が，動詞が v^E，その目的語が名詞 n^E である動詞句であるとする．このとき，日本語文書において，$C(n^J, v^J \mid 『を』)$ が大きいならば，英語文書において，$C(n^E, v^E \mid \mathbf{obj})$ も大きい傾向にある．

「ベクトル空間を張る」を例にして提案手法の基本アイデアを述べる．「ベクトル空間を張る」の翻訳として適切な『張る』の訳語を v^E とする．本手法の適用は，翻訳対象である日本語コロケーション「n^J を v^J」の共起性 $C(n^J, v^J \mid 『を』)$ が比較的大きいことを前提とする．このとき，仮定 2より

(A) $\Delta = \{V^E : C(trans(n^J), V^E \mid \mathbf{obj}) \geq \theta_E\}$ とおくと，$v^E \in \Delta$である．

ここで $trans(n^J)$ は n^J の英訳を表わす．ただし，日本語名詞の英訳は曖昧さなく求まると仮定する．たとえば，

$trans(『ベクトル空間』)$ = “vector space”

である．また，ここで，閾値 θ_E を導入し，C が θ_E 以上であるとき，その共起性が大きいと判断する．

日本語文書において$C(N^J, 『張る』 \mid 『を』)$の値が大きい名詞 N^J としては，『蜘蛛の巣』，『罠』，『空間』，『テント』，『ネットワーク』，『頬』などがある．このうち，「頬を張る」の『張る』は「ベクトル空間を張る」の『張る』とは異なる意味で用いられているが，「蜘蛛の巣を張る」，「罠を張る」，「空間を張る」，「テントを張る」，「ネットワークを張る」の『張る』はどれも「ベクトル空間を張る」の『張る』とほぼ同じ意味で用いられている．これらの名詞の集合

$\Gamma_J =$
{『蜘蛛の巣』, 『罠』, 『空間』, 『テント』, 『ネットワーク』}

に対して，

$\Gamma_E = \{N^E : N^E = trans(N^J),\ N^J \in \Gamma_J\}$
$=$ {“web”, “trap”, “space”, “tent”, “network”}

を用意する．このとき 仮定 1より，

(B) 次の集合

$\{N^J \in \Gamma_J :$「N^J を張る」の適切な翻訳が，動詞が v^E，その目的語が N^E である動詞句，$N^E = trans(N^J)\}$

の要素数は大きい（$|\Gamma_J|$ に近い）傾向にある．

さらに，仮定 2より

(C) 「N^J を張る」($N^J \in \Gamma_J$) の適切な翻訳が，動詞が v^E，その目的語が N^E である動詞句（た

†2 本稿では名詞の訳語が一意で動詞の訳語の候補が複数ある場合について取り扱っているが，中には動詞の訳語が一意で名詞の訳語の候補が複数あるという場合も考えられる．その場合は訳語が一意である方の単語（動詞）を手掛かりにして，もう一方の単語（名詞）の訳語を求めることになる．

だし $N^E = trans(N^J)$) であるならば，共起性 $C(N^E, v^E \mid \mathbf{obj})$ が大きい．

(A), (B), (C)より，「ベクトル空間を張る」の『張る』に対する適切な英訳 v^E は，$V^E \in \Delta$ のうちで

$$\left\{N^E \in \Gamma_E \ : \ C(N^E, V^E \mid \mathbf{obj}) \geq \theta_E\right\}$$

の要素数が大きいものだと言える．

2.3 アルゴリズム

前節のアイデアを定式化し，「n^J を v^J」の翻訳として優先度付で v^J の訳語候補を求めるアルゴリズムを示す．ただし，前述したように，提案手法は，$C(n^J, v^J \mid$『を』$)$ が比較的大きいコロケーション「n^J を v^J」を翻訳対象とする．

1. 日本語名詞 N_i^J について，共起性 $C(N_i^J, v^J \mid$『を』$)$ の大きいものから順に N_i^J を利用者に提示する．利用者は提示された日本語コロケーション $\langle N_i^J, $『を』$, v^J\rangle$ における v^J の意味が $\langle n^J, $『を』$, v^J\rangle$ における v^J の意味とほぼ同一かどうかを調べ，同一だと思われる N_i^J を Γ_J の要素に加える．これを $|\Gamma_J| = m$ になるまで行なう．
2. Γ_J 内の各日本語名詞に対応する以下の英語名詞の集合 Γ_E を求める．

   $\Gamma_E = \left\{N^E : N^E = trans(N^J), N^J \in \Gamma_J\right\}$.

3. n^E を n^J の訳語とし，以下の英語動詞の集合 Δ を求める．

   $\Delta = \{V^E \ : \ C(n^E, V^E \mid \mathbf{obj}) \geq \theta_E\}$.

4. 各 V^E ($\in \Delta$) に対し，以下の評価値 $E(V^E)$

   $E(V^E) = \left|\{N^E \in \Gamma_E : C(N^E, V^E \mid \mathbf{obj}) \geq \theta_E\}\right|$

   を与え，これを優先度（高い方を優先）とし，$E(V^E)$ の高いものから順に $(V^E, E(V^E))$ を出力する．

なお，θ_E は実験的に定める閾値である．

3. 実　　装

3.1 システムの概要

提案手法を用いて，コロケーション翻訳支援システムを実装する．本システムは英語文書作成の際の支援システムとして，一般利用者に自由に利用してもらうことを目標におく．また，本手法で抽出するWWW上での共起情報を保存しておけば，それらは有用な言語資源として他の自然言語技術への活用も期待できる．そのために，本システムは，PerlによるCGIプログラムとしての実装を採用し，webコンテンツとしてwebブラウザから利用することにする．webコンテンツとして実装することにより，webブラウザでwebコンテンツを閲覧できる環境さえあれば誰でも利用でき，また，共起データの再利用性も高まる．

Fig. 1 Image of CGI application(1).

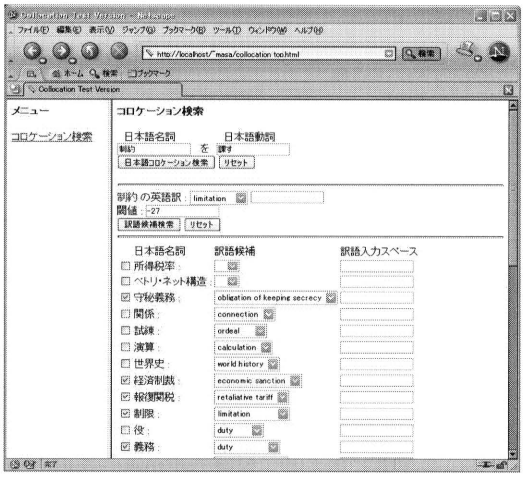

Fig. 2 Image of CGI application(2).

実装したコロケーション翻訳支援システムを実装した各ステップでの動作状況を**Fig. 1**–**Fig. 3** に示す．利用者は，まず翻訳対象の日本語コロケーションを成す日本語名詞 n^J と日本語動詞 v^J とを入力し「日本語コロケーション検索」ボタンを押す（**Fig. 1**）．すると，システムは v^J と共起する日本語名詞 N_i^J を共起性の大きい順番に並べて提示する（**Fig. 2**）．このとき，システムが表示する日本語名詞リストの一部を**Fig. 4** に示す．システムは，左から順に，提示した日本語名詞をΓ_Jの要素として採用するかどうかを選択するチェックボックス，v^Jと共起する日本語名詞 N^J，対訳単語辞書から得られる N^J の訳語候補，N^J の訳語を入力する入力スペース，を表示する．

利用者は，提示された N^J と v^J とのコロケーション

Fig. 3　Image of CGI application(3).

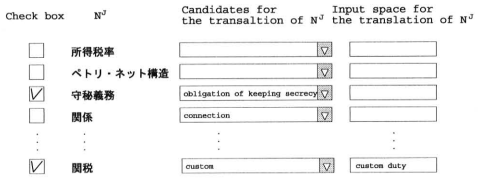

Fig. 4　List of Japanese nouns which co-occur strongly with v^J.

$\langle N^J, 『を』, v^J \rangle$ を考え，そのときの v^J が，翻訳対象「n^J を v^J」の v^J とほぼ同じ意味で用いられているかどうかを確認する．もし「n^J を v^J」の v^J と同じ意味の v^J と共起するものならば，チェックボックスで選択して N^J を Γ_J の要素に加える．候補の中から，このような名詞を適当数選択して Γ_J を作成する．母語話者である日本人にとって，日本語コロケーションにおける動詞の意味は容易に同定することができ，この作業に対する負担は小さい．

また，日本語名詞（n^J および各 N^J）に対する訳語は，日本語名詞と一緒に提示された訳語候補のうちから選択する，あるいは利用者が訳語を直接入力することのどちらかで設定され，これによって Γ_E が作成される．閾値 θ_Eについては，実験的に適切であると思われた-27 をデフォルト値として与える．ただし，適切な θ_E は利用者の要求によって変わってくる．θ_Eを大きくすれば，システムが提示する訳語候補の数が増え，訳語としての妥当性のチェックに要する利用者の負担は大きくなるが，妥当な訳語が取りこぼされる可能性は低くなる．逆にθ_Eを小さく設定すれば，システムが出力する訳語候補の妥当性チェックに対する利用者の負担は軽減されるが，妥当な訳語が取りこぼされる可能性が高くなる．利用者は自分の要求に合わせて θ_E を自由に設定することもできる．

このようにして，日本語名詞とその訳語，および閾値 θ_E に対する設定が終了したら，最後に「訳語候補検索」ボタンを押してシステムの回答を待つ．システムはWWW上から v^J の訳語候補を抽出し，評価値 E の高い順に並べて利用者に提示する（**Fig. 3**）．

利用者は，提示された訳語候補から妥当だと思われる訳語を選択するわけだが，システムが提示した訳語 V^E を，

(1)　意味的に妥当だと思われる，
(2)　明らかに意味的に妥当でない，
(3)　どちらか分からない，

の三つに 分別することは容易に行なえる．このうち，(3)については改めて辞書で調べるなどして妥当な訳語かどうかを判断すればよい．各訳語候補には，実際にV^Eとn^Eとから成る英語コロケーションが用いられているwebページへのハイパーリンクも共に提示され，実文書での用例を見ながら訳語の妥当性を判断するのに利用できる．

3.2　日本語コロケーションの抽出

2.3節のアルゴリズムの 1. で，翻訳対象「n^J を v^J」に対して，日本語動詞 v^J と強く共起する日本語名詞を抽出するのには，EDR日本語コーパス(JCO-V020E)のデータを用いる．日本語の共起性もEDR日本語コーパス上の共起性で評価する．Γ_J の作成については網羅性はあまり重要ではなく，v^J の訳語候補を抽出する手掛かりとして，v^J と強く共起する名詞をいくつか用意できれば十分なため，既存のコーパスから算出される$C(N^J, v^J \mid 『を』)$ を基にΓ_J の候補を求める．EDRコーパスから抽出したΓ_J の候補は$C(N^J, v^J \mid 『を』)$の大きな順に並べられて，利用者に提示される．

ただし，EDRコーパス中で見つかる共起の組み合わせの中には専門用語のような一般に使用頻度の低い名詞が含まれることがある．コロケーションが出現頻度の低い単語を含む場合，相互情報量による共起性評価は信頼性が低くなる．また，このような低頻度の名詞については，利用者が単語の意味，あるいはその単語に対する訳語を知らない場合が多く，リスト内に候補として提示しても Γ_J の要素に採用されることは少ないと思われる．そこで，名詞の出現頻度が低い名詞（具体的には，コーパス中に10回未満しか出現しなかった名詞）については，予め候補から除いておくものとする．

3.3　日本語名詞に対する訳語

利用者が入力したn^Jおよび前節で作成したΓ_Jの候補N^Jは，提示する際に単語対訳辞書で訳語候補を調べ，一緒に提示する．$trans(N^J)$ は，用意された訳語候補の中から選ぶか，あるいは利用者が直接入力することで設定される．日本語名詞に対して訳語候補を求めるのにはEDR日英対訳辞書（JEB-V016）を用いる．日英対訳辞書内に N^J に対応する訳語が存在する場合には，その訳

語候補をリストとして表示する．利用者が Γ_J の要素として選択した日本語名詞 N^E に対して，エディットボックスに訳語が入力されていた場合にはその訳語を，入力されていなかった場合には，訳語候補リストから選択された訳語を$trans(N^E)$ として採用し，これによって Γ_E を作成する．

3.4 WWW からの Δ の候補の抽出

前節で求めた $n^E = trans(n^J)$ に対してweb上から v^E の訳語候補を次のように求める．web上の文書からデータを抽出するのには web検索エンジン（AltaVista[†3] を使用）を利用する．検索エンジンでは検索キーαを入力し検索を行なうと，次の情報を得ることができる．

ヒット数(hit count) ：検索キー α を含むページの数．

抜粋(extract) ：検索キー α を含むページへのURLとそのページの一部（検索キー α を含む部分）．

Δの要素となり得る動詞候補は，検索エンジンの検索結果ページの抜粋部分から抽出する[†4]．まず，n^E を検索キーとして検索エンジンに掛け，その検索結果を求める．結果ページの抜粋を文単位に分け，抜粋内の文のうち，n^E を含む文をすべて抜き出す．抜き出した文を品詞taggerで品詞付けし，文中の動詞を抽出する．ここで，品詞付けにはTreeTagger [†5] を用いる．文内のすべての動詞をΔの要素の候補として収集する．

3.5 英語共起性の計算

Δ の各要素 V^E と n^E との共起性を求め，評価値 E を計算する．V^E と n^E との共起性や E は検索エンジンの検索結果のヒット数を利用して計算する．$\widetilde{C}(n^E, V^E \mid \mathbf{obj}) \geq \theta_E$ なるV^EをΔ の要素とする．ここで，$\widetilde{C}$ については後述する．

共起性 $C(N^E, V^E \mid \mathbf{obj})$ は，以下のように表わされる．

$$\log \frac{\dfrac{f(\langle N^E, \mathbf{obj}, V^E\rangle)}{K(\mathbf{obj})}}{\dfrac{f(\langle N^E, \mathbf{obj}, *\rangle)}{K(\mathbf{obj})} \cdot \dfrac{f(\langle *, \mathbf{obj}, V^E\rangle)}{K(\mathbf{obj})}}.$$

$f(\langle N^E, \mathbf{obj}, V^E\rangle)$ は$\langle N^E, \mathbf{obj}, V^E\rangle$ の WWW 上の英語文書全体での頻度であり，また，

$$f(\langle N^E, \mathbf{obj}, *\rangle) = \sum_{V^E} f(\langle N^E, \mathbf{obj}, V^E\rangle),$$
$$f(\langle *, \mathbf{obj}, V^E\rangle) = \sum_{N^E} f(\langle N^E, \mathbf{obj}, V^E\rangle)$$

である．$K(\mathbf{obj})$ は関係$\mathbf{obj}$ でのWWW上の文書内でのコロケーションの総数で，

$$K(\mathbf{obj}) = \sum_{N^E}\sum_{V^E} f(\langle N^E, \mathbf{obj}, V^E\rangle)$$

である．これらの値を既存の検索エンジンを用いて求めるのであるが，用いた検索エンジン AltaVista の使用上の制約から，すべての英語文書をダウンロードすることはできず，上記の値を正確に求めることはできない．そこで，今回は，検索エンジンが出力するヒット数を用い，以下のような近似を行なう．

$$f(\langle N^E, \mathbf{obj}, V^E\rangle) \simeq h(\text{“}V^E \text{ the } N^E\text{”}) + h(\text{“}V^E \text{ a } N^E\text{”}) \quad (2)$$
$$f(\langle N^E, \mathbf{obj}, *\rangle) \simeq h(N^E) \quad (3)$$
$$f(\langle *, \mathbf{obj}, V^E\rangle) \simeq h(V^E) \quad (4)$$

$h(\alpha)$ は，α を検索キーとして検索した際に，検索エンジンから得られるヒット数である．

$C(N^E, V^E \mid \mathbf{obj})$ はこの近似を用いて

$$C(N^E, V^E \mid \mathbf{obj}) \simeq \log \frac{h(\text{“}V^E \text{ the } N^E\text{”}) + h(\text{“}V^E \text{ a } N^E\text{”})}{h(N^E)\cdot h(V^E)} + \log K(\mathbf{obj}) \quad (5)$$

のように表わされる．$h(\alpha)$ は α の出現頻度ではなく，α を含む Web ページの数であること，および，形態素解析・構文解析を施すわけではないことから，(5)は誤差を含む．しかし，今回の実験では，これを第一次近似として用いた．また，$K(\mathbf{obj})$ も検索エンジンでは求まらないため，$C(\langle N^E, \mathbf{obj}, V^E\rangle)$ の代わりに，

$$\widetilde{C}(N^E, V^E \mid \mathbf{obj}) = \log \frac{h(\text{“}V^E \text{ the } N^E\text{”}) + h(\text{“}V^E \text{ a } N^E\text{”})}{h(N^E)\cdot h(V^E)}. \quad (6)$$

を用いた．上記(2), (3), (4)の近似が正しいとしても，$\widetilde{C}$ は $\log K(\mathbf{obj})$ だけ実際の$C(N^E, V^E \mid \mathbf{obj})$ よりも小さな値となるが，この問題は閾値 θ_E を低く設定することで回避できる．ただし，θ_E の見積もりは使用する検索エンジンの持つデータ量に依存するため，実験的にしか求めることはできない．

3.6 E の計算と結果の表示

システムはまず n^E と Δ の各要素 V^E との共起性を求める．このとき求めた各 $V^E \in \Delta$ との共起性はファイ

†3 http://www.altavista.com/

†4 AltaVista の検索結果ページでは，ヒット数が 1,000 以上あっても，1,000 ページ分の URL(とその抜粋) までしか提示されず，それ以上のページについては辿ることはできない．しかし，動詞候補を得るのには 1,000 ページ分の抜粋で十分だと考え，ここから動詞候補を抽出することとする．

†5 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

ルとして保存しておき，後に再び n^E に関する共起性を求める場合には，この共起データを再利用する．次に $\widetilde{C}(n^E, V^E \mid \mathbf{obj}) \geq \theta_E$ であったV^Eについて，検索エンジンを使って各$N^E \in \Gamma_E$との共起性を求める．結果は n^E の場合と同様に再利用できるようにファイルに保存しておく．

n^E との共起が θ_E 以上の動詞 V^E について，$E(V^E)$ を計算し，訳語候補 V^E を評価値 E の大きい順に並べて表示する．また，共起性の計算に用いるときに検索した，検索キー "V^E the n^E" での検索結果の抜粋部分から実文書へのURLを抽出しておき，訳語候補と共にそのURL（すなわち，"V^E the n^E"が用いられている実文書へのハイパーリンク）も表示し，訳語の妥当性のチェックの際に利用してもらう．

4. おわりに

本稿では，web上での単語の共起性に基づくコロケーション翻訳支援システムの実装について述べた．本システムは，翻訳対象のコロケーションを成す動詞 v^J と名詞 n^J を入力し，システムが提示する日本語名詞候補を選択することで，半自動的にコロケーション「n^J を v^J」におけるv^Jに対する訳語候補をweb文書から抽出し利用者に提示する．

現在のところ，Δ の候補，および各訳語候補の共起性を求める度に検索エンジンにアクセスするため，一回の検索に時間が掛かるという問題がある．実用的なシステムを構築するためには，共起に関する情報を予めローカルに保存しておく必要がある．また，係り受け解析などを行なって，共起情報を予め求めておくことができれば，より高い精度での訳語抽出が期待できる．

WWWは潤沢な文書資源ではあるものの，基本的にあらゆる人が自由に情報を発信できるため，必ずしも各言語の表現の自然さと言う点で良質なデータばかりではない．これに対して，文書が，母語話者が書くような良質な文書であるか否かを判別することも試みている[4]．これによって，比較的質の高いと思われる文書のみを取り出すことができれば，本手法の精度はよりよいものになると期待される．

参考文献

1) A. Kilgarriff and G. Greffenstette: Introduction to the Special Issue on Web as Corpus, *Information Technology Research Institute, University of Brighton, also published in Computational Linguistics*, **29.3, 1–15, 2003**.
2) M. Shibata, Y. Tomiura and S.Tanaka: A Mehod for Retrieving Translations for Collocation in Web Data, *Proc. of Asian Symposium on Natural Language Processing to Overcome Language Barriers*, **1, 1–8, 2004**.
3) D. Hindle: Noun Classification from Predicate-Argument Structure, *Proc. of ACL*, **28, 268–275, 1990**.
4) 藤井 宏，田中 省作，冨浦 洋一: Skew Divergence に基づく母語話者/非母語話者文書の判別，**FIT2004 情報科学技術レターズ**, **81–85, 2004**.