Web上の語の共起性に基づいたコロケーションの翻訳支援

柴田雅博[†,☆] 冨浦洋一[††] 田中省作[†††,☆☆]

母語以外の言語で文を書く際には，辞書や実文書に載っている例文を調べるなどして，その目的言語の文として不自然なコロケーション（語と語の組合せ）を用いないように気をつけなければならない．しかし，辞書に載っている例文は数が少なく，辞書を調べただけでは妥当な訳語候補を求めるのは難しい．一方，実文書を調べて妥当な訳語候補を求めるのには大変な労力が必要である．本稿では，Web上の文書から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する．本稿では特に日本語の動詞 v^J と目的語 n^J とからなるコロケーション「n^J を v^J」を英語の動詞 v^E と目的語 n^E とからなるコロケーションに翻訳することを対象とする．ただし，n^J の妥当な訳語が n^E であることはすでに分かっているものとする．提案手法では，Web上の文書における単語の共起性を元に，「n^J を v^J」における v^J の妥当な訳語 v^E の候補を半自動的に獲得する．また，評価実験により，提案手法を用いて，高い精度で v^J に対する妥当な訳語候補が得られることを示す．最後に，本手法を動詞と目的語のコロケーション以外のパターンに適用することについて検討する．

Assisting with Translating Collocations Based on the Word Co-occurrence on the Web Texts

Masahiro Shibata,[†,☆] Yoichi Tomiura[††] and Shosaku Tanaka[†††,☆☆]

Non-native speakers have to consult dictionaries or real documents in order not to use unnatural collocations (combinations of words). However, it is difficult to seek out the proper candidates for the translation of a collocation using dictionaries, because there is a few examples in them. On the other hand, a lot of efforts are needed to get the proper translation using real documents. This paper proposes a new method for retrieving candidates for the proper translation of a collocation from web texts. The method can be performed for any source and target languages, but in this paper we focus on the case of translating a Japanese phrase "n^J WO v^J" into an English phrase, which consists of a predicate verb v^E and its object n^E, under the condition that n^E is already known as the proper translation of n^J. The proposed method seeks candidates for v^E (v^E is the translation of v^J) based on co-occurrence in web texts. The experimental result shows that the proposed method can seek out the proper candidates for v^E with high reliability. Finally, this paper discusses the extension of the proposed method to deal with the other types of collocations.

1. まえがき

母語以外の言語を的確に運用するためには，文法や個々の単語の意味といった知識のほかに，自然な語と語の組合せ（コロケーション☆☆☆）に関する知識が欠かせない．非母語話者が書いた文書の中で，文法的には妥当であるのに不自然に見える表現には，不自然なコロケーションが含まれている場合が多い．たとえば，日本語で「犯罪を行う」というのは，意味は通じるかもしれないが不自然な表現である．語彙知識の乏しい

† 九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
†† 九州大学大学院システム情報科学研究院
Graduate School of Information Science and Electrical Engineering, Kyushu University
††† 九州大学情報基盤センター
Computing and Communications Center, Kyushu University
☆ 現在，九州システム情報技術研究所
Presently with Institute of Systems & Information Technologies/KYUSHU
☆☆ 現在，立命館大学文学部人文科学総合インスティテュート
Presently with Institute of Human Science, College of Letters, Ritsumeikan University

☆☆☆ 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' の組合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．

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，辞書や実際の目的言語で書かれた文書に頼るほかないが，それを正確に判断するのは難しい．

例として，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，どれが適切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うのには相当な労力と時間が必要となる．

一方，近年インターネットの爆発的普及により Web 上には膨大な言語データが蓄積され，現在も日々増加し続けている．この Web 上の文書を 1 つの用例集 “Web as corpus” と見なし[1]，有効活用する試みが行われている．また，Web 文書はその量もさることながら，言語の多様性という点でも有用であり，言語に依存しない手法であれば，多言語への適用可能性も高くなる．

そこで，本稿では母語以外での文書作成時のコロケーション翻訳支援を目的として，Web から半自動的にコロケーションの訳語候補を抽出する手法を提案する．提案手法は，言語処理知識や計算機知識に長けていない一般利用者でも利用できるシステムを想定する．提案手法では，単語の共起性を利用して各訳語候補の順位付けを行い，利用者に提示する．その際，訳語候補のコロケーションが用いられている例文を Web 文書から抽出して提示することもできるので，利用者が行う訳語候補の妥当性のチェック作業の効率化につながる．

なお，本稿では翻訳における原言語として日本語，目的言語として英語を想定する．また，2 章から 4 章までは，日本語の動詞 v^J と名詞 n^J が「を」格でつな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 からなるコロケーションに翻訳する場合に限定して議論する．このように限定したのは，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．それ以外のパターンのコロケーションについては，5 章で考察する．

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 2) 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 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 2 つの仮定に基づいている☆．

[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 $\langle n_i^J, 『を』, v^J \rangle$ $(i = 1, 2, \cdots, k)$ における v^J の意味が同一ならば，それらの日本語

☆ 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，なか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．

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」を例にして本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．

提案手法は，共起性 $C(n^J, v^J \mid 『を』)$ が比較的大きいコロケーション $\langle n^J, 『を』, v^J \rangle$ を翻訳対象とする．したがって，仮定 2 より

(A) $\Delta = \left\{ V^E : C(trans(n^J), V^E \mid \mathbf{obj}) \geq \theta_E \right\}$ 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．すなわち，

$trans(『ベクトル空間』) =$ “vector space”

である．また，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，

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

$= \{$“web”, “trap”, “space”, “tent”, “network”$\}$

を用意する．このとき仮定 1 より，

(B) 次の集合

$\left\{ N^J \in \Gamma_J :$「$N^J$ を張る」の適切な翻訳が，v^E を動詞，$trans(N^J)$ をその目的語とする動詞句 $\right\}$

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，v^E を動詞，$trans(N^J)$ をその目的語とする動詞句であるならば，共起性 $C(trans(N^J), v^E \mid \mathbf{obj})$ 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いえる．

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きで 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，$C(N_i^J, v^J \mid 『を』)$ 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 $\langle N_i^J, 『を』, v^J \rangle$ 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 $|\Gamma_J| = m$ になるまで行う．なお，後で述べる実験では，経験的に $m = 10$ 個程度とした．

(2) Γ_J の各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) $n^E = trans(n^J)$ とし，以下の英語動詞の集合 Δ を求める．

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

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

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

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，閾値 θ_E は実験的に定める．

3. 実　　装

アルゴリズムの (1) で日本語の共起性を求めるのには EDR 日本語コーパス（JCO-V020E）を用いる．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 との共起性の大きな名詞がいくつかあれば十分なため，今回は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 を作成する．また，Web 検索エンジン（実験では AltaVista☆を使用）の検索結果から，得られる次の情報を用いて，Δ や $E(V^E)$ を求める．

ヒット数（hit count）：検索キー α を含むページの数．

☆ http://www.altavista.com/

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

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

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）．

以下，その詳細について述べる．

3.1 Web からの Δ の候補の抽出

Δ の要素となりうる動詞候補は，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☆．まず，n^E を検索キーとして検索エンジンにかけ，その検索結果を求める．結果ページの抜粋内で n^E を含む文を抜き出し，その文内のすべての動詞を Δ の要素の候補として収集する．品詞付けには TreeTagger ☆☆を用いる．その後，n^E と動詞候補 V^E との共起性を求め，$\widetilde{C}(n^E, V^E \mid \mathbf{obj}) \geq \theta_E$ なる V^E を Δ の要素する．$\widetilde{C}$ については次節で述べる．

3.2 英語共起性の計算

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

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

$f(\langle N^E, \mathbf{obj}, V^E\rangle)$ は $\langle N^E, \mathbf{obj}, V^E\rangle$ の Web 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 **obj** での Web 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式 (5) のように表される．$h(\alpha)$ 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，式 (5) は誤差を含む．しかし，今回の実験では，これを第 1 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，式 (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．

4. 実　　験

提案手法は，辞書だけでは妥当な訳語が判断できないコロケーションの翻訳に特に有効だと思われる．したがって，提案手法の有効性について正確に評価するためには，たとえば辞書の例文に記載されていないコロケーションのみを対象とした評価実験を行うのが最も直接的である．しかし，この場合，システムが出力する訳語候補が妥当かどうかをどう判断するかが問題となり，たとえば人間の内省で判断するなどすれば評価が客観的なものとならない．

そこで，本稿では，まず辞書の例文からは妥当な訳語が見つからなかったいくつかの日本語コロケーションについて，提案手法を用いて実際に訳語候補を求めた場合のシステムの出力結果を例示する．次に，提案手法の定量的な評価のために，テストコロケーション

☆ 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たどることはできない．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．

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

☆☆☆ 相互情報量は，正の相関があるとき正の値を，負の相関があるとき負の値をとるが，$\widetilde{C}$ は $\log K(\mathbf{obj})$ を引いた値のため，正の相関があっても負の値をとる場合がある．

表 1 ベクトル空間（vector space）を張る
Table 1 Result of “BEKUTORU-KUKAN WO HARU”.

動詞	E	名詞									
		罠	蜘蛛の巣	枝	キャンプ	コネクション	リンク	ネットワーク	根	空間	テント
		trap	web	branch	camp	connection	link	network	root	space	tent
create	8	−24.7	−23.7	−25.0	−26.6	−23.8	−24.1	−24.1	−24.6	−23.7	−25.8
use	7	−24.5	−24.1	−26.0	−26.2	−25.6	−23.4	−24.7	−25.0	−23.7	−24.6
define	6	−24.1	−26.4	−24.7	$-\infty$	−24.4	$-\infty$	−24.6	−23.2	−24.1	−27.2
construct	5	−24.0	−24.8	$-\infty$	−25.0	−25.1	$-\infty$	$-\infty$	$-\infty$	−24.6	−23.7
include	4	−25.2	−25.0	−25.4	−25.5	−25.7	−23.9	−24.9	−25.3	−23.8	−24.8
like	4	−23.7	−25.5	−25.2	−25.3	−26.0	−26.1	−25.6	−24.6	−23.4	−23.7

表 2 勝利（victory）を決める
Table 2 Result of “SHOURI WO KIMERU”.

動詞	E	名詞				
		ゲーム	試合	ゴール	得点	優勝
		game	match	goal	point	championship
win	4	−22.0	−22.6	−25.7	−24.9	−21.4
match	4	−23.9	−23.1	−24.6	−23.7	−25.3
set	3	−25.6	−24.5	−22.0	−23.7	$-\infty$
score	3	$-\infty$	−24.2	−22.2	−22.5	$-\infty$
point	3	−24.7	−24.9	$-\infty$	−23.4	$-\infty$
bowl	3	−21.2	−24.5	$-\infty$	$-\infty$	−20.3
race	2	−23.6	−26.6	$-\infty$	$-\infty$	−22.5
get	2	−24.4	−25.4	−25.9	−23.6	−26.1
earn	2	−25.8	$-\infty$	−27.4	−21.8	−24.1

を辞書の例文から作成し，提案手法によるシステムが出力する r 位タイ以内の訳語候補に正解が含まれる割合（正解率）を求める．訳語候補が正解であるか否かの判定は，客観性を重視し，4.2.3 項で述べるように複数の辞書の例文に基づいて行う．ここでテストコロケーションの作成に辞書の例文を利用したのは，各テストコロケーションに対して少なくとも 1 つは信頼できる正解を用意しておくためである．本手法では Γ^J から Γ^E を求める際に対訳辞書を用いているものの，辞書の例文を訳語抽出の足がかりとするわけではない．また，辞書の例文に載っていないコロケーションは辞書に載っているコロケーションと比べて発生頻度は小さいが，共起性 $\widetilde{C}$ の点では辞書に載っているものと辞書に載っていないものとに，統計的に大きな差は生じない．実際，辞書に載っていないコロケーションと辞書に載っているコロケーションについて，共起性 $\widetilde{C}$ を比べてみたところ，ともに平均 −23 程度で大きな差は認められなかった．そのため，本手法で辞書に載っているコロケーションの訳語を求めることと辞書に載っていないコロケーションの訳語を求めることとに，大きな性能差は現れないと考えられる．

4.1 訳語候補抽出例

まずはじめに，辞書を調べても妥当な訳語が見つからなかったいくつかの日本語コロケーションに対して，提案手法で得られた訳語候補の抽出例を**表 1**，**表 2**，**表 3** に提示する．これらは $\theta_E = -25$ と設定した場合の各訳語候補とその評価値 E を示したものである．各表において，第 2 行の日本語名詞は用意した Γ_J の要素，第 3 行の英語名詞はその訳語である．また，第 1 列は訳語候補，第 2 列はその評価値 E である．表内の数字は各コロケーションに対する $\widetilde{C}$ である．たとえば，表 1 において “construct” と “web” との共起性は

$$\widetilde{C}(\text{“web”}, \text{“construct”} \mid \mathbf{obj}) = -24.8$$

となっている．また，“construct” の評価値は

$$E(\text{“construct”}) = 5$$

である．検索エンジンで $\langle N^E, \mathbf{obj}, V^E \rangle$ が見つからなかった場合には $\widetilde{C}$ は $-\infty$ としている．

これらの表を見ると，辞書では妥当な訳語が見つからなかったコロケーションについても，提案手法で訳語を求めると，v^J の訳語候補として妥当だと思われる英語動詞の評価値 E が比較的高くなっていることが分かる．

得られた各訳語候補を，(a)「n^J を v^J」に対する v^J の訳語として妥当そうに見える動詞，(b) 意味的に明らかに違っている動詞，(c) 判定の難しい動詞に判

表 3 スケジュール（schedule）を組む
Table 3 Result of “SUKEJURU WO KUMU”.

動詞	E	名詞				
		プログラム	ローン	組織	予算	プラン
		program	loan	organization	budget	plan
establish	5	−23.2	−24.0	−24.5	−23.1	−23.9
plan	4	−24.5	−24.6	−25.9	−23.7	−23.3
manage	4	−24.1	−24.7	−24.6	−23.2	−25.5
create	4	−24.1	−26.0	−24.9	−24.1	−24.3
complete	4	−23.5	−23.3	−26.4	−24.9	−24.9
approve	4	−23.9	−21.6	−25.6	−21.0	−22.1
work	3	−23.7	−27.0	−24.8	−26.2	−23.1
study	3	−22.5	−24.9	−25.5	−25.5	−23.8
select	3	−23.4	−24.1	−25.1	−26.2	−24.4
review	3	−24.0	−25.2	$-\infty$	−24.2	−24.2
prepare	3	−24.8	$-\infty$	−25.7	−22.4	−23.1
pay	3	−24.9	−23.1	−27.5	−26.0	−23.6
develop	3	−22.6	$-\infty$	−25.2	−23.4	−22.0

別することは，人間が見れば容易である．このとき，判定の難しいものや，意味を知らない動詞については，改めて辞書を調べたり，システムが候補とともに抽出した Web 上の実文書と照らし合わせて確認したりすればよい．また，上記で妥当そうだと判断した動詞が本当に妥当かどうかを確認する場合にも同様の作業を行えばよい．数個程度の動詞であれば，この作業に対する利用者への負担は小さい．

4.2 評価実験

次に，提案手法の有効性を定量的に評価するために行った小規模な評価実験について述べる☆1．

4.2.1 テストコロケーションの作成

テストコロケーションの作成は，まずランダムに選んだ v^E について，英和辞書☆2で v^E を引き，その項に例文が載っていればその例文（日本語文）から日本語コロケーション $\langle n^J, 『を』, v^J \rangle$ を抜き出し，これをテストデータとする☆3．実験では 41 個の日本語コロケーションをテストセットとして用意した．用意したテストセットについては，その一覧を付録に載せる．

4.2.2 Γ_J の設定

実際に得られる動詞候補は，Γ_J で選択される名詞にも依存する．名詞数が少ないと動詞の訳語候補に対して十分な順位付けができず，また名詞数が多すぎると $\langle n^J, 『を』, v^J \rangle$ の v^J と同一語義であることを保つのが難しくなり，ノイズとしてしか働かない不適切な名詞が多く含まれる可能性が高くなる．そこで今回は経験的に EDR コーパスで共起性の大きかったものから順に人手で $\langle n^J, 『を』, v^J \rangle$ の v^J と同一語義であるものをチェックし，$|\Gamma_J|$ が 10 程度（平均 10.2 個）となるように名詞集合を用意した．

4.2.3 正解か否かの判定法

ある $\langle n^J, 『を』, v^J \rangle$ の v^J の訳語として妥当な英語動詞をすべて列挙するのは困難である．本稿では訳語が妥当かどうかの判定をできるだけ客観的に行えるように，コーパスその他の情報を用いた判断を行わず，辞書（複数を用意☆4）の例文としてあがっているかどうかを，正解の一次近似として扱い，正否を判定することとする☆5．

具体的には $\langle n^J, 『を』, v^J \rangle$ の v^J の訳語として，以下の (1)〜(4) の場合の V^E を正解と見なす．

(1) V^E を（テストコロケーションの作成に用いた英和辞書も含め）いくつかの英和辞書で調べた場合に，V^E の例文として英語コロケーション $\langle n^E, \mathbf{obj}, V^E \rangle$，その和訳として $\langle n^J, 『を』, v^J \rangle$ あるいはこれとほぼ同義の日本語コロケーションが見つかった場合．

(2) n^J の訳語 n^E をいくつかの英和辞書で調べた場合に，n^E の例文として英語コロケーション $\langle n^E, \mathbf{obj}, V^E \rangle$，その和訳として $\langle n^J, 『を』, v^J \rangle$ あるいはこれとほぼ同義の日本語コロケーションが見つかった場合．

(3) v^J をいくつかの和英辞書で調べた場合に，v^J

☆1 実験で用いた Web データは 2004 年 2 月時点のものである．
☆2 新英和中辞典（第 7 版）（研究社，2003）を使用した．
☆3 英和辞書を用いたのは和英辞書よりも載っている例文が豊富だったためである．
☆4 使用した辞書は，新英和中辞典（第 7 版）（研究社，2003），ジーニアス英和辞典（初版）（大修館，1994），新和英中辞典（第 5 版）（研究社，2002），カレッジライトハウス和英辞典（初版）（研究社，1995），英辞郎（アルク，2002）である．
☆5 本実験で失敗と判定されたもののうち，用いる辞書を増やしたりネイティブ・チェックを行ったりすれば，成功だと判定されるものも存在する可能性がある．

の例文として $\langle n^J, 『を』, v^J \rangle$ あるいはこれとほぼ同義の日本語コロケーション，その英訳として $\langle n^E, \mathbf{obj}, V^E \rangle$ が見つかった場合．

(4) n^J をいくつかの和英辞書で調べた場合に，n^J の例文として $\langle n^J, 『を』, v^J \rangle$ あるいはこれとほぼ同義の日本語コロケーション，その英訳として $\langle n^E, \mathbf{obj}, V^E \rangle$ が見つかった場合．

たとえば，$\langle 『注意』, 『を』, 『引く』 \rangle$ の『引く』の訳語として，“capture”，“attract”，“draw”，“win”，“arrest”，“pull” が妥当か否かを判定する場合を考える．まず，各英語動詞を英和辞書で調べたところ，このうちの “capture” の例文から $\langle 『注意』, 『を』, 『引く』 \rangle$ が見つかりその訳文から $\langle$“attention”, **obj**, “capture”$\rangle$ が見つかったとする．このとき，(1) より “capture” は妥当な訳語である．また，“attention” を英和辞書で調べたところ，例文から $\langle 『注意』, 『を』, 『引く』 \rangle$ が見つかりその訳文から $\langle$“attention”, **obj**, “attract”$\rangle$ と $\langle$“attention”, **obj**, “draw”$\rangle$ が見つかったとする．このとき，(2) より “attract” と “draw” は妥当な訳語である．次に，『引く』を和英辞書で調べたところ，例文から $\langle 『注意』, 『を』, 『引く』 \rangle$ が見つかりその訳文から $\langle$“attention”, **obj**, “win”$\rangle$ が見つかったとする．このとき，(3) より “win” は妥当な訳語である．最後に，『注意』を和英辞書で調べたところ，例文から $\langle 『注意』, 『を』, 『引く』 \rangle$ が見つかりその訳文から $\langle$“attention”, **obj**, “arrest”$\rangle$ が見つかったとする．このとき，(4) より “arrest” は妥当な訳語である．“pull” は，(1)〜(4) のいずれにもあてはまらなかったため，妥当な訳語ではないと見なす．

4.2.4 実験結果とその評価

4.2.1 項のテストコロケーションセットに対して提案手法を用いて訳語候補を求めた場合に，評価値 E の上位 r 位タイ以内の候補に正解が含まれるテストコロケーションの割合を求めた．

提案手法で訳語候補を求めたとき，$r = 1, \cdots, 5$ および $r = 10, 20$ として閾値 θ_E を変えたときの正解率を**図 1** に示す．図 1 を見ると，5 位以内としたときは $\theta_E = -23$ のとき最も高く 0.80，10 位以内としたときは $\theta_E = -23$ および -24 のとき最も高く 0.83，20 位以内としたときは $\theta_E = -24$ のとき最も高く 0.92 であった．提案手法は単語の共起性だけを用いており，シソーラスなど意味的な制約を入れていないにもかかわらず，非常に高い割合で妥当な訳語を求めることができることが分かる．

訳語候補集合 Δ には次のような要求がある．

- Δ 中に高い確率で少なくとも 1 つは正解が含まれる．
- Δ 中から正解を見つけるのに利用者が調べなければならない訳語候補数が小さい．

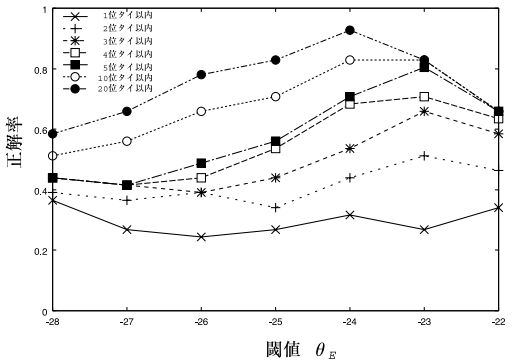


図 1 Web から訳語候補を抽出したときのテストセットに対する正解率
Fig. 1 Accuracy for test set using the web.

θ_E を低くすれば，提案手法で求まる訳語候補数は多くなるため，Δ 中に少なくとも 1 つは正解が含まれる確率は大きいが，利用者は正解を見つけるために多くの訳語候補を調べる必要がある．逆に，θ_E を高くすれば，提案手法で求まる訳語候補数は少なくなるため，利用者は少しの訳語候補しか調べなくて済むが，Δ 中に正解が含まれる確率は小さい．

そこで，θ_E を変えたとき，Δ 中に少なくとも 1 つは正解が含まれる割合と正解を見つけるまでに利用者が調べなければならない訳語候補数との関係を調査した．ここで，正解を見つけるまでに利用者が調べなければならない訳語候補数を，Δ 中の訳語候補を評価値 E の高い順に見ていき，正解を見つけるまでに調べた訳語候補数☆とした．ただし，Δ 中に 1 つも正解が見つからなかったものについては，調べた候補数を $|\Delta|$ とした．その調査結果を**図 2** に示す．Δ 中に少なくとも 1 つは正解が含まれる割合と利用者が調べなければならない訳語候補数とは，相反する関係になる．θ_E の設定は，2 つのうち，どちらかを優先させるかに依存する．

次に，訳語候補の抽出に Web 文書を用いることの有効性について考える．本稿では，和英辞書を用いても妥当な訳語が見つからない場合にも対処できるように，訳語候補 Δ の抽出に Web 文書を用いている．この有効性を示すために，辞書を用いて訳語候補を求めた場合と，Web を用いて訳語候補を求めた場合とで，妥当な訳語がどのくらい求まるのかを調査した．つま

☆ 最初に見つかった正解の評価値 E が E_0 であり，E_0 以上の評価値を持つ候補が n 個ある場合には，n 個まで調べるものとして計算した．

表 4 Web からの訳語抽出と辞書からの訳語抽出
Table 4 Extraction from the web and extraction from a Japanese-English dictionary.

		S_2	
		正解を含む（個）	正解を含まない（個）
S_1	正解を含む（個）	32	8
	正解を含まない（個）	0	1

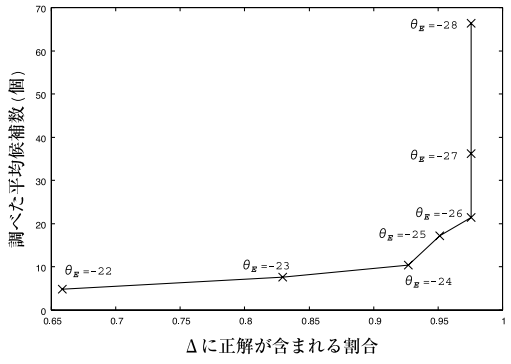


図 2 閾値 θ_E の設定
Fig. 2 Setting of threshold θ_E.

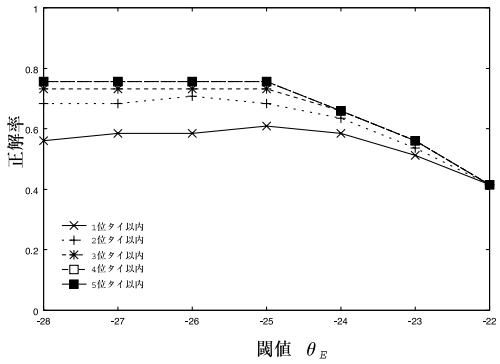


図 3 辞書から訳語候補を抽出したときのテストセットに対する正解率
Fig. 3 Accuracy for test set using a dictionary.

り，4.2.1 項で作成したテストセット中の各日本語コロケーション $\langle n^J, 『を』, v^J \rangle$ に対して，

(1) Web 上で n^E との共起が見つかったすべての動詞の集合 S_1，
(2) 和英辞書で v^J の訳語から抽出した訳語候補集合 S_2，

とし，S_1，S_2 の各要素に対して，4.2.3 項で示した方法で候補集合が正解を含むかどうかを調べた．S_2 は，和英辞書（カレッジライトハウス和英辞典（初版））で v^J を引いたときに，項目 v^J 内であげられているすべての訳語候補を抜き出し，作成した．その結果を**表 4** に示す．これを見ると，テストセット 41 個中，辞書からは 9 個のコロケーションについて妥当な訳語を得ることができなかったのに対し，Web を用いた場合に妥当な訳語を得ることができなかったのは 1 個だけであった．また，Web で妥当な訳語が得られなかった 1 個については，辞書でも訳語を得ることができなかった．ただし，得られた訳語候補数を見ると，Web から得られる平均訳語候補数が 430 個であったのに対し，辞書から得られる平均訳語候補数は 5.8 個と非常に少なく，辞書から訳語を求める方が絞り込みの点では有利である．

辞書から求めた訳語候補に対して，提案手法で訳語候補の絞り込みを行った場合，$r = 1, \cdots, 5$ として閾値 θ_E を変えたときの正解率を**図 3** に示す．図 1 と図 3 を比較すると，r が 1，2 と小さい場合には辞書から訳語を抽出した方が正解率が高いものの，r を大きくすると，Web から訳語を抽出した方が正解率が高くなることが分かる．

5. 考　　察

5.1 Γ_J と訳語候補との関係

提案手法で得られる訳語候補は，用意する Γ_J に依存する．どの名詞を Γ_J の要素に加えるかは，利用者の判断に委ねられることとなる．しかし，本手法は共起の全体的な傾向を見ているだけなので，Γ_J（Γ_E）に不適切な名詞（利用者は「n^J を v^J」と同一語義と判断したが，実際は同一語義とはいえないもの）が少数含まれていたとしても，その名詞と共起する動詞の傾向は，多くの場合，他とは大きく異なっていると考えられ，評価値 E に対する多少のノイズとなるものの，おしなべて見れば全体の評価値への影響は少ない．たとえば，表 1 において「テントを張る」の「張る」は設置するという行為を表していると考えるならば「ベクトル空間を張る」の「張る」とは異なるともいえるが，仮に「テント」を除いたとしても “create” と “define” 以外の動詞の E の値が 1 つずつ下がるだけで，全体の順位には大きな影響は与えない．したがって，Γ_J を作成する際の語義判定に対して利用者に厳密な判定を要求するものではない．

また，実際の利用を考えると，満足のいく訳語候補が得られなかった場合には Γ_J を変更して（語義の絞り込みを厳しく，あるいは緩くして）もう一度検索を

行うといった，インタラクティブな利用も考えられる．

5.2 Web 文書を用いる利点と欠点

本稿では Web 文書を訳語選択の言語資源として採用したが，もちろん Web 文書の代わりにオーソライズされたコーパスを使っても，提案手法で妥当な訳語を抽出することは可能である．しかし，共起性は高いが構成要素の語の 1 つあるいはすべてが低頻出語であるようなコロケーションの場合は，オーソライズされたコーパスでは共起データが見つからないというデータ・スパースネスの問題がある．本手法では $\langle n^J, 『を』, v^J \rangle$ の v^J の訳語候補を $n^E = trans(n^J)$ と強く共起する動詞の中から求めるため，コーパス内に n^E との共起が見つからなければ失敗する．そのため，v^J の訳語候補集合 Δ を求める際に，極力，候補の中から正解の v^E がもれることのないように，Web 文書を使用する．

ただし，Web 文書には次のような欠点もある．まず，オーソライズされた文書ではないため，Web 上から得られるデータは良質なものだけではない．たとえば，文書が非母語話者によって書かれたものならば，文法的，語用的誤りを含む可能性がある．ほかにも，日々更新されるため再現性がないこと，Web 検索エンジンを用いていることから正確な共起性が計算できないこと，ネットワークへのアクセス回数が大きいため実行時間がかかることなどがある．

5.3 他のコロケーションパターンへの対応

これまでは，「n^J を v^J」という日本語コロケーションパターンを，動詞が v^E，その目的語が n^E という英語コロケーションパターンに翻訳することに限定して議論してきた．ほかにも「n^J が v^J」という日本語コロケーションパターンから英語の主語 n^E と動詞 v^E とからなる英語コロケーションパターンへの翻訳ならば，本手法を単純に適用可能である．本節では，上記以外のパターンの翻訳について，提案手法の適用可能性を議論する．もちろん，本節であげた少数の例をもって，あらゆるパターンの翻訳における本手法の有効性を実証できるわけではないが，将来，共起情報を正しく抜き出せるようになったときのための試金石にはなると思われる．

「n^J を v^J」という日本語コロケーションパターンの翻訳は，つねに動詞が v^E，その目的語が n^E という英語コロケーションパターンであるとは限らない．ほかにも，ある前置詞 p^E をともなって，p^E に目的語 n^E が係って前置詞句を成し，その前置詞句が v^E に係るという翻訳パターンも考えられる．また，日本語コロケーションパターン「n^J に v^J」の場合も同様である．翻訳対象となりうるパターンは他にも考えられるが，英語側のコロケーションが動詞 v^E，前置詞 p^E，前置詞の目的語 n^E からなるパターンの場合ならば，既存の技術でも次のようにして一応の対処が可能である．関係 f が前置詞 p^E であった場合の英語コロケーションの共起性は，

$$C(N^E, V^E \mid p^E) = \log \frac{f(\langle N^E, p^E, V^E \rangle) \cdot K(p^E)}{f(\langle N^E, p^E, * \rangle) \cdot f(\langle *, p^E, V^E \rangle)} \qquad (7)$$

で表される．ここで $K(p^E)$ は Web 上の全文書における関係 p^E での英語コロケーションの総数である．また，$f(\langle N^E, p^E, V^E \rangle)$，$f(\langle N^E, p^E, * \rangle)$，$f(\langle *, p^E, V^E \rangle)$ はそれぞれ，検索エンジンのヒット数を用いて次のように近似される．

$$\begin{aligned}
&f(\langle N^E, p^E, V^E \rangle) \simeq \\
&\quad h(\text{“}V^E\ p^E \text{ the } N^E\text{”}) + h(\text{“}V^E\ p^E \text{ a } N^E\text{”}) \\
&\quad + h(\text{“}V^E * p^E \text{ the } N^E\text{”}) \\
&\quad + h(\text{“}V^E * p^E \text{ a } N^E\text{”}) \\
&\quad + h(\text{“}V^E * * p^E \text{ the } N^E\text{”}) \\
&\quad + h(\text{“}V^E * * p^E \text{ a } N^E\text{”}), \\
&f(\langle N^E, p^E, * \rangle) \simeq \\
&\quad h(\text{“}p^E \text{ the } N^E\text{”}) + h(\text{“}p^E \text{ a } N^E\text{”}), \\
&f(\langle *, p^E, V^E \rangle) \simeq \\
&\quad h(\text{“}V^E\ p^E\text{”}) + h(\text{“}V^E * p^E\text{”}) \\
&\quad + h(\text{“}V^E * * p^E\text{”}).
\end{aligned}$$

ここで検索キー中の “ * ” は，AltaVista におけるワイルドカードであり，任意の 1 単語とマッチするものである☆．$\widetilde{C}(N^E, V^E \mid p^E)$ はこれらの近似を用いて計算される．

また，Δ の抽出については次のように変更すればよい．あらかじめ英語における関係 f の集合を用意する．前述の場合と同様に，検索結果の抜粋から n^E と共起する動詞 V^E を抜き出した後，すべての関係 f との組合せ (V^E, f) に対して，n^E の間の共起性 $\widetilde{C}(n^E, V^E \mid f)$ を求め，θ_E 以上になった (V^E, f) を Δ の要素とする．

上記の拡張を行ったシステムを用いて，「経験を生かす」（“draw on an experience”）に対する訳語候補を求めてみた．Γ_J として，「教訓」（“lesson”），「長所」（“virtue”），「力」（“power”），「テーマ」（“theme”），「リズム」（“rhythm”），「特徴」（“characteristic”），「資格」（“license”），「資源」（“resource”）を用意した．その結

☆ 2004/4/1 より AltaVista の仕様が変更された．それにともない，現在（2004/10/8 時点）はワイルドカードを使用できない．

果,(draw, on) に対する評価値 E の値は $\theta_E = -19$ のときに 7 であり,これは関係 “on” におけるコロケーションの中で第 1 位であった.

また,別の例として「栄誉に輝く」(“cover (one-self) with glory”) に対する英訳候補を求めた.Γ_J として,「喜び」(“pleasure”),「勝利」(“win”),「栄冠」(“coronal”),「名誉」(“honor”),「トップ」(“top”),「三冠王」(“triple crown”) を用意した.その結果,(cover, with) に対する評価値 E の値は $\theta_E = -19$ のときに 4 であり,これは関係 “with” におけるコロケーションの中で第 2 位であった.

これら 2 つの例を見ても,本手法を「n^J を v^J」から動詞が v^E,その目的語が n^E のパターン以外への翻訳にも適用可能であることが分かる.

ただし,E による順序付けは同一の f の中でしか意味を持たない.それは,共起性 $\widetilde{C}$ には,関係 f によって異なるバイアス ($-\log K(f)$) が与えられており,単純比較ができないためである.関係 f によらず全候補を一様に E で比較するのであれば,共起性を $\widetilde{C}$ ではなく C で評価する必要がある.もし,あらかじめ共起に関する情報をローカルに蓄えておくことができるならば,C を求めることができ,上記の問題は解決される.実際にこれを行うには,容量の面で問題が生じると思われるかもしれないが,Γ_J (Γ_E) の要素とするのは辞書に載っている基本名詞で十分であり,すべての名詞や名詞句をカバーする必要はないため,実現は十分可能である.

また,動詞と目的語(あるいは主語と動詞)のようなパターンとは違い,動詞と前置詞句の関係を求めるためには離れた関係を扱う必要があり,パターンマッチでは検索精度が落ちるため,本手法の精度も下がると予測される.もし,係り受けを含む検索クエリを扱える検索エンジンが利用可能ならば,あるいは上記のように共起情報をローカルに蓄えておくことができれば,もっと柔軟なコロケーションパターンに対しても取り扱うことができるものと期待できる.

なお,提案手法で妥当な訳語が求まる可能性があるのは,

- 日本語コロケーションの共起性が大きい,
- 日本語コロケーションを翻訳したときに,どのような英語コロケーションパターンとなるのか,その候補が分かっている,

場合に限られる.

5.4 訳語の曖昧性

本稿では日本語名詞 N^J に対する英語の訳語 $trans(N^J)$ が曖昧さなく求まることを前提としたが,一般に,ある単語に対する訳語は複数存在するため,本来は Γ_J から Γ_E を作成する際に,妥当な訳語を絞り込む必要がある.しかし,本手法は共起の全体的な傾向を見るものであるから,少数のノイズ(すなわち N^E が「N^J を v^J」に対する N^J の妥当な翻訳となっていない場合)を含んでも,精度はそれほど下がらないと予測される.これについては今後検討していく予定である.

6. 関連研究

文献 3), 4) では,名詞句に対する翻訳を対象に,コロケーションに対する訳語候補を,検索エンジンのヒット数を用いて評価している.しかし,そのコロケーションに対する訳語候補は,句を構成する各単語を辞書引きしたときに得られた訳語の組合せによって作成しており,前述のような辞書に載っていない訳語には対処できない.それに対し本手法は訳語候補自体も Web 上の文書データから抽出するため,このような場合にも対処できる.文献 5) では,Web ではなく BNC (British National Corpus) におけるコロケーションの発生確率で訳語を評価する.また,係りの種類を考慮していること,削除補間によってデータ・スパースネスの問題にも対処していることなどが上記 2 編と異なる.しかし,コロケーションの訳語候補は辞書で求めた訳語の組合せで作成しており,やはり辞書に載っていない訳語には対処できない.

文献 6) では,本稿と同じく動詞句(特に動詞と目的語の組合せ)に対する翻訳を対象としている.ただし,こちらは動詞の訳語が確定している場合に目的語の妥当な訳語を求めるのが目的である.文献 6) では,名詞をサブカテゴリ化した,クラス情報付き格フレーム辞書を用意し,学習データから推定した動詞 v とクラス c の組 (c, v) に対する n の発生頻度 $f(n|c, v)$ に基づいて,訳語候補の絞り込みを行う.この手法は機械翻訳分野ではある程度成功が見込まれるが,本稿での目的にはそぐわない.まず,本稿の目的を考えると,利用者はどんな $\langle n^J, 『を』, v^J \rangle$ を入力するのかが分からないため,用意する格フレーム辞書の網羅性が問題となる.また,クラスの粒度をどの程度に設定すればいいのかを調整するのも難しい.さらに,これもやはり名詞の訳語候補は,文献 3)〜5) と同様に辞書から抜き出したものに限定されており,やはり辞書に載っていない訳語には対処できない.

文献 7), 8) では,単語の共起性を用いて妥当な訳語の絞り込みを行っている.これらの手法は日英で対応のとれた文書(コンパラブル・コーパス)をソースと

表 5 実験で使用したテストセット
Table 5 Test set used in the examination.

n^J	を	v^J
穴	を	開ける
蚊帳	を	張る
注意	を	引く
塩	を	入れる
ハンドル	を	切る
灰	を	撒く
権利	を	守る
嘘	を	見破る
美しさ	を	残す
場所	を	見つける
眠り	を	妨げる
瓶の栓	を	抜く
戦い	を	挑む
語彙	を	増やす
名誉	を	保つ
希望	を	持つ
宿題	を	出す
信用	を	固める
戒厳令	を	敷く
試合	を	中止する
経験	を	積む
方程式	を	解く
癇癪	を	堪える
スキャンダル	を	公表する
評判	を	落とす
階段	を	登る
時計	を	調節する
ペンキ	を	塗る
紛争	を	解決する
手	を	組む
辞書	を	調べる
グラス	を	空ける
目	を	そらす
声明	を	まとめる
議長	を	任命する
物語	を	話す
勘定	を	払う
休暇	を	すごす
敬意	を	払う
生計	を	稼ぐ
電気	を	伝える

し，同内容の文書には意味の同じ単語が多く用いられているという仮定の下，文献 7) では対応する英日文書 (D_E, D_J) の各単語 t_E，t_J（$t_E \in D_E$，$t_J \in D_J$）の共起性から，文献 8) では訳語対応の取れた英中単語対を手がかりに訳語を知りたい単語と共起する語の傾向（ベクトル空間モデル）から，未知語に対する訳語を求める．しかし，これらはあくまで未知語の訳語対応を求めるものであって，係り受け関係に応じて訳語が変わるものは想定していない．それに対して，本研究では係り受け関係に応じて訳語が変わることを想定している．

7. むすび

本稿では，コロケーションに対する訳語候補を Web 上の文書から抽出する手法を提案し，評価実験とその結果を示した．

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たびに検索エンジンにアクセスする．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あらかじ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って，共起情報をあらかじ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．

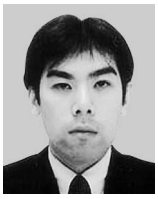
付録

4.2 節で訳語を求めたテストセットを**表 5** に提示する．

参考文献

1) Kilgarriff, A. and Grefensette, G.: Introduction to the Special Issue on Web as Corpus, Technical Report ITRI-03-20 (2003).
2) Hindle, D.: Noun Classification from Predicate-Argument Structure, *Proc. 28th ACL*, pp.268–275 (1990).
3) 池野 篤，村田稔樹，下畑さより，山本秀樹：インターネット自然言語資源を利用した機械翻訳，沖テクニカルレビュー 182 号，Vol.67, No.1, pp.49–52 (2000).
4) Grefensette, G.: The World Wide Web as a Resource for Example-based Machine Translation Tasks, *Translating and the Computer 21: ASLIB'99*, London, UK (1999).
5) Tanaka, T. and Baldwin, T.: Translation Selection for Japanese-English Noun-Noun Compounds, *Machine Translation Summit IX*, pp.378–385 (2003).
6) Prescher, D., Riezler, S. and Rooth, M.: Using a Probablistic Class-Based Lexicon for Lexical Ambiguity Resolution, *18th International Conference on Computational Linguistics*, Vol.19, No.3, pp.331–358 (2000).
7) 堀内貴司，千葉靖伸，浜本 武，宇津呂武仁：言語横断検索により自動収集された日英関連報道記事からの訳語対応の獲得，情報処理学会研究報告，2002-NL-150, pp.191–198 (2002).
8) Fung, P. and Yee, L.Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 36th ACL and 17th COLING*, pp.414–420 (1998).

(平成 16 年 10 月 18 日受付)
(平成 17 年 4 月 1 日採録)



柴田 雅博（正会員）

1996 年九州大学工学部情報工学科卒業．2005 年同大学大学院システム情報科学府博士後期課程学位取得退学．現在，九州システム情報技術研究所特別研究助手．自然言語処理，身体障害者支援に関する研究に従事．電子情報通信学会，言語処理学会各会員．

冨浦 洋一（正会員）

1984 年九州大学工学部電子工学科卒業．1989 年同大学大学院工学研究科博士課程単位取得退学．同年九州大学工学部助手，1995 年同助教授，1996 年同大学大学院システム情報科学研究科助教授，2000 年同大学院システム情報科学研究院助教授，現在に至る．工学博士．自然言語処理，計算言語学，人工知能に関する研究に従事．言語処理学会，人工知能学会各会員．

田中 省作（正会員）

2000 年九州大学大学院システム情報科学研究科博士後期課程修了．同大学情報基盤センター，高等研究機構助手を経て，現在，立命館大学文学部助教授．博士（工学）．自然言語処理，言語処理技術の外国語教育への応用に関する研究に従事．言語処理学会，英語コーパス学会，日本ケルト協会各会員．