WWW ドキュメントからの日本語共起に対する英訳候補の検索

冨浦 洋一 *，　柴田雅博 **，　田中省作 ***

* 九州大学 大学院システム情報科学研究院
** 九州大学 大学院システム情報科学府
*** 九州大学 情報基盤センター

{tom,shibata}@lang.is.kyushu-u.ac.jp,　sho@cc.kyushu-u.ac.jp

1 はじめに

日本語単語の英訳は和英辞書を用いて得ることができるが，英訳が複数ある場合，そのうち，その単語が使われている環境（文脈，前後関係）で適切な英訳はどれかを，辞書に記載されている用例を頼りに求める必要がある．しかし，辞書の性質上，十分な用例は記載されていないため，一般にはこの選択は難しい．さらに，得られた英訳がどれも適切でない場合さえある．たとえば,「ベクトル空間を張る」の英訳を考えよう．和英辞書で『張る』の英訳を調べてみと，“stretch”，“pitch”，“stick”，“extend”，“cover” などが得られるが，適切な英訳はこれらに含まれていない.『張る』の類義語の英訳を調べることも考えられるが，やはり，適切なものがどれかを用例を頼りに判断するのは困難である．さらに，類義語の範囲を広げすぎると，そもそも「ベクトル空間を張る」の意味を保存しなくなる可能性もある．

適切な英訳を求めるには，辞書を活用しながら，実際の英文書を調べて見るのが確実である．しかし，これを手作業でやるとすると，相当の労力と時間を要す．一方，WWW 上の英語文書集合は，大規模な用例集と考えることができる．この中から，日本語単語の英訳の候補とその使用例を抽出することができるならば，適切な英訳を求めるための作業の効率化に繋がる．

本稿では，対象を「n^J を v^J」[1] に限定して，WWW 上の英語文書の共起情報に基づく「n^J を v^J」の英訳として適切な v^J の英訳の候補を絞り込む手法を提案する．v^J の英訳候補 v^E が十分な精度で絞り込めていれば，“v^E a n^E” （n^E は n^J の英訳）などを検索キーとして求まる使用例を，適切な英訳を判断するための情報として活用できる．

[1] n^J は日本語名詞，v^J は日本語動詞．また，n^J の英訳は既知であるか辞書などで容易に分かるものとする．

2 共起情報を利用した英訳候補の絞り込み

提案手法は,「n^J を v^J」の適切な英訳が，動詞を v^E，その目的語を名詞 n^E とする自動詞句であることを想定している．まず，基本的なアイデアを述べたあと，v^J の英訳候補を絞り込むための具体的なアルゴリズムについて述べる．

2.1 Basic Idea

本稿では，語 w が関係（あるいは関係を規定する機能語）f を介して語 w' に係る構造を共起と呼び，$\langle w, f, w' \rangle$ と記す．また，関係 f で結ばれる場合の w と w' の相関の程度を $\langle w, f, w' \rangle$ の共起性と呼び，$C(w, f, w')$ と記す．共起性としては，文献 [1] などで用いられている相互情報量に基づく値

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

を想定している．ここで，$P(\langle w, f, w' \rangle)$ は関係が f であるときの共起 $\langle w, f, w' \rangle$ の条件付発生確率であり，

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

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

である．つまり，$C(w, f, w')$ は f を介した共起における 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．

提案手法は，共起性 $C(n^J, \text{『を』}, v^J)$ が比較的高い共起「n^J を v^J」の英訳として適切な v^J の英訳を求めることを前提とする．

ここで，以下の 2 つの仮定を置く．

仮定 1 日本語名詞 n_1^J, n_2^J, $\cdots$, n_k^J に対して，共起「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，それらの日本語共起の英訳として適切な v^J の英訳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，動詞を v^E，その目的語を名詞 n^E とする動詞句であるとする．このとき，日本語文書において，$C(n^J, 『を』, v^J)$ が高いならば，英語文書において，$C(n^E, \mathbf{obj}, v^E)$ も高い傾向にある．

「ベクトル空間を張る」を例に考えてみる.「ベクトル空間を張る」の英訳として適切な『張る』の英訳を v^E とする．日本語文書において，$C(N^J, 『を』, 『張る』)$ が高い名詞として,『蜘蛛の巣』,『罠』,『空間』,『頬』などがある.「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」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．したがって,「ベクトル空間を張る」,「蜘蛛の巣を張る」,「罠を張る」,「空間を張る」の適切な英訳は，“v^E a vector space”，“v^E a web”，“v^E a trap”，“v^E a space” であると考えられ（仮定 1），v^E の候補は，共起性 $C($“vector space”$, \mathbf{obj}, V^E)$ がある程度高いもののうち，多くの $N^E \in \{$“web”, “trap”, “space”, $\cdots\}$ に対して $C(N^E, \mathbf{obj}, V^E)$ がある程度高い V^E であると考えられる（仮定 2）．

2.2 Algorithm

前述のアイデアに基づいた,「n^J を v^J」の英訳として適切な v^J の英訳の（優先度付の）候補を求めるアルゴリズムを示す．ただし，日本語名詞の英訳は曖昧さなく求まるものと仮定する．

1. 以下の日本語名詞の集合 Γ'_J を求める．

$$\Gamma'_J = \{N^J \mid C(N^J, 『を』, v^J) \geq \theta_J\}.$$

2. ユーザの手作業により，以下の Γ_J を求める．

$$\Gamma_J = \{N^J \in \Gamma'_J \mid \text{共起 } \langle N^J, 『を』, v^J \rangle \text{ における } v^J \text{ の意味が } \langle n^J, 『を』, v^J \rangle \text{ における } v^J \text{ の意味とほぼ同一．}\}.$$

3. 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（電子化版対訳辞書を使用）．

$$\Gamma_E = \{N^E \mid N^E \text{ は } N^J (\in \Gamma_J) \text{ の英訳 }\}.$$

4. n^E を n^J の英訳とし，以下の英語動詞の集合 Δ を求める．

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

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

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

を与え，これを優先度（高い方を優先）として，$(V^E, E(V^E))$ を出力する．

θ_J，θ_E は実験的に定めるスレッシュホールドである．ただし，本手法は，共起性 $C(n^J, 『を』, v^J)$ が比較的高いものに対して適用できる手法であるため，スレッシュホールド θ_J は，

$$C(n^J, 『を』, v^J) \geq \theta_J$$

となるように設定しなければならない．さらに，これを満たすように θ_J を設定すると，非常に小さな値になるような場合は，おそらく，上記 4 で求まる Δ 中に v^J の適切な英訳は含まれないと思われる（つまり提案手法が適用できない）．

3 実験

日本語共起「n^J を v^J」の英訳として適切な v^J の英訳候補を求める実験を行なった．ただし，極小規模な実験であり，定量的な評価には至っていない．

3.1 共起性を求める手順

今回の実験では，日本語共起に関しては，WWW 上の日本語文書ではなく，EDR 電子化辞書の日本語コーパス [2] を使用して求めた．

英語共起に関しては，WWW 上の英語文書から求めた．共起性 $C(N^E, \mathbf{obj}, V^E)$ は，以下のように表せる．

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

ただし，$f(\langle N^E, \mathbf{obj}, V^E \rangle)$ は共起 $\langle N^E, \mathbf{obj}, V^E \rangle$ の WWW 上の英文書全体での頻度であり，

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

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

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ないため，上記の値を正確に求めることはできない．そこで，今回の実験では，検索エンジンの出力するヒット数を用い，以下のような近似を行なった．

$$f(\langle N^E, \mathbf{obj}, N^E \rangle) \simeq h(\text{“}V^E \text{ the } N^E\text{”}) + h(\text{“}V^E \text{ a } N^E\text{”}) \quad (1)$$
$$f(\langle N^E, \mathbf{obj}, * \rangle) \simeq h(N^E) \quad (2)$$
$$f(\langle *, \mathbf{obj}, V^E \rangle) \simeq h(V^E) \quad (3)$$

$h(\alpha)$ は α を検索キーとした場合の検索エンジンが出力するヒット数である．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・構文解析を施しているわけではないことから来る様々な誤差を含む．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 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，

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

を用いた．上記 (1)(2)(3) の近似が正しいとしても，$\widetilde{C}$ は $\log K$ だけ実際の $C(\langle N^E, \mathbf{obj}, V^E \rangle)$ よりも小さな値となるが，この問題はスレッシュホールドの設定の仕方で回避できる．

Γ_E を求める際，N^J（$\in \Gamma_J$）に対して，複数の英訳が存在する場合は,(勿論，複数選択しても構わないが,) そのうち任意に一つを選んだ．また，Δ を求める際，全ての英語動詞 V^E に対して，$C(n^E, \mathbf{obj}, v^E)$ を求めるのは困難であるため，n^E を検索キーとしたときの検索結果（検索キーを含む部分の抜粋，最大 1000 件）から，tree tagger による品詞付与と簡単なパターンマッチにより，n^E を目的語とする動詞集合を求め，この集合の部分集合として Δ を求めた．

3.2 実験結果

本稿で例として取り上げた「ベクトル空間を張る」の他,「条件を飲む」,「訴訟を起こす」,「最適値を求める」,「わさびをおろす」それぞれに対して行なった結果を表 1〜5 に示す．表中の名詞は，2.2 節の Γ_J および Γ_E であり，動詞 v の行中の名詞 n の欄の数値は $\widetilde{C}(n, \mathbf{obj}, v)$ である．また，表中の動詞は，$\widetilde{C}$ に対する閾値 θ_E を -25 とし，評価値 $E(V^E)$ が上位のものを挙げている．表のキャプションの括弧中に，適切と考えられる英訳を記している．表より，適切な英訳の動詞に対する評価値 E が比較的高いものであることが分かる．

また，研究社新英和中辞典中の用例より「《他動詞》+《名詞句》」の形態の英語自動詞句とその日本語訳（「《名詞》+を+《動詞》」）の組 22 個をランダムに選択し，日本語訳の動詞の英訳候補を抽出する実験を行なった．用例に示された通りの英語動詞を候補の上位 5 位までに含む割合は 36%，他の辞書や抽出された使用例などから正しい英訳であると判断できるものを候補の上位 5 位までに含む割合は 86% であった．

4 おわりに

WWW 上の英語文書を利用した，日本語共起「n^J を v^J」の英訳として適切な v^J の英訳の候補を絞り込む手法を提案した．提案手法の特徴は，英語文書における共起情報に基づいて，v^J の英訳候補の妥当性を数量化することにある．提案手法は，v^J そのものの対訳情報や日本語や英語のシソーラスなどを使用していないにもかかわらず，かなりの精度が得られている．v^J の英訳候補が対訳辞書から得られる場合は，より確からしいと考えられる．したがって，候補の提示法として，本手法で提案した評価値に従った提示だけでなく，補足情報として，対訳辞書で得られる v^J の英訳か否かの情報も付与することも考えられる．

また，本稿では，日本語名詞の英訳は曖昧さなく求めることができることを前提としているが，この制約をはずしても同様の手法で日本語共起の英訳を求めることができると考えられる．

参考文献

[1] Hindle, D.: Noun Classification from Predicate-Argument Structures, *Proc. 28th Annual Meeting of ACL*, pp. 268–275 (1990).

[2] EDR 電子化辞書 日本語コーパス (JCO-V020E).

表 1: 「ベクトル空間を張る」({construct, define, create} a vector space)

動詞	E	『を』格の日本語名詞（その対訳）									
		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: 「条件を飲む」(accept a condition)

動詞	E	『を』格の日本語名詞（その対訳）				
		話	異義	涙	要求	依頼
		speech	objection	tear	demand	request
consider	3	-25.6	-22.7	-27.5	-25.0	-23.6
satisfy	3	$-\infty$	-23.8	$-\infty$	-21.2	-22.0
present	3	-24.8	-24.3	-27.2	-24.7	-25.7
evaluate	3	-24.2	$-\infty$	-27.1	-24.8	-24.0
state	2	-24.4	-25.0	-28.1	-25.8	-25.6
examine	2	-25.5	-24.1	$-\infty$	-24.7	-25.1
accept	2	-26.5	-24.3	-27.6	$-\infty$	-22.9
create	2	-25.5	-26.8	-26.3	-23.6	-24.8
determine	2	-26.6	-24.8	-27.3	-24.5	-27.2
meet	2	-27.2	-23.8	-28.1	-20.9	-25.6
permit	2	$-\infty$	-24.7	-27.2	-26.1	-23.3

表 3: 「訴訟を起こす」({bring, file} a suit)

動詞	E	『を』格の日本語名詞（その対訳）			
		行動	裁判	事業	革命
		action	trial	business	revolution
bring	2	-23.5	-26.6	-25.2	-24.9
claim	2	-23.9	-24.3	-26.5	-26.9
maintain	2	-24.4	-26.8	-23.5	-27.6
separate	2	$-\infty$	-23.0	-24.3	$-\infty$
show	2	-25.6	-24.1	-23.9	-27.3
authorize	1	-23.6	$-\infty$	$-\infty$	$-\infty$
follow	1	-24.9	-25.7	-26.9	-27.1
.	.	.	.	.	.
:	:	:	:	:	:
threaten	1	-24.0	$-\infty$	$-\infty$	-25.5
file	0	-25.2	-26.8	-27.3	$-\infty$

表 4: 「最適値を求める」({determine, calculate} an optimum value)

動詞	E	『を』格の日本語名詞（その対訳）							
		値	経路	値段	合計	逆行列	現在時刻	大きさ	素数
		value	route	price	total	inverse matrix	current time	size	prime-number
determine	8	-21.9	-24.2	-24.2	-23.4	-22.6	-23.5	-23.1	-22.9
calculate	7	-21.8	$-\infty$	-23.4	-21.5	-19.3	-22.8	-23.4	-21.0
define	6	-23.6	-24.3	-26.1	-25.0	-23.2	-24.0	-23.8	-22.7
find	6	-24.3	-24.8	-24.1	-25.7	-21.9	-24.7	-26.0	-22.3
generate	5	-23.9	-24.8	-26.3	-24.7	-22.8	-25.6	$-\infty$	-21.0
obtain	5	-23.6	-25.7	-24.4	$-\infty$	-21.7	-22.9	-26.3	-23.6
contain	5	-22.6	-25.9	-26.4	-24.9	-24.1	-24.6	-26.6	-23.8
specify	5	-22.6	-24.1	$-\infty$	-24.6	$-\infty$	-24.1	-22.5	-25.4
select	5	-24.1	-24.4	-25.0	-26.3	$-\infty$	-24.6	-22.5	-24.4

表 5: 「わさびをおろす」(grate a horseradish)

動詞	E	『を』格の日本語名詞（その対訳）								
		大根	山芋	生姜	ポテト	りんご	タマネギ	ニンニク	チーズ	レモンの皮
		Japanese-radish	yam	ginger	potato	apple	onion	garlic	cheese	lemon-peel
add	9	-24.1	-24.6	-22.3	-23.7	-24.3	-20.7	-20.7	-22.4	-21.2
cut	9	-21.6	-23.7	-24.4	-23.1	-23.6	-22.1	-23.3	-22.8	-23.0
grate	9	-17.5	-22.5	-20.2	-20.9	-21.7	-20.1	-21.9	-19.2	-17.9
like	9	-24.1	-24.3	-23.7	-22.1	-23.0	-22.9	-23.3	-23.1	-24.7
chop	8	$-\infty$	-23.1	-21.2	-22.2	-22.0	-18.4	-19.6	-23.5	-21.0
combine	8	$-\infty$	-24.1	-22.7	-22.8	-22.7	-21.1	-21.1	-22.0	-20.6
mix	8	$-\infty$	-24.0	-23.7	-23.5	-22.8	-22.6	-22.4	-22.8	-22.5
slice	8	$-\infty$	-22.1	-21.4	-22.0	-21.4	-19.5	-21.4	-21.1	-21.0