Examination about Usefulness of the Method for Translating Collocations Based on the Word Co-occurrence on the Web Documents

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/10.15017/1516060

出版情報：九州大学大学院システム情報科学紀要. 10 (1), pp.45-49, 2005-03-25. Faculty of Information Science and Electrical Engineering, Kyushu University
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

コロケーション翻訳支援システムに対する有用性の調査

柴田 雅博* ・ 冨浦 洋一** ・ 田中 省作***

Examination about Usefulness of the Method 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 examines usefulness of the method for translating collocations (combinations of words) based on the word co-occurrence on the web documents by practical writing test of English documents. 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. Besides, it is difficult to seek out the proper candidates for the translation of a collocation even using dictionaries, because they don't have enough examples. This paper makes an examination of usefulness of the proposed method in writing English documents practically, and tries to show the proposed method to be effective for seeking out the proper translation.

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

1. はじめに

本稿では，web文書における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[1)2)]の有用性についての調査を行なう．

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，自然な語と語との組み合わせ（コロケーション†1）に関する知識が欠かせない．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．

これに対し，web文書中での単語の共起性に基づき，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，この手法を用いたコロケーション翻訳支援システムを構築している[1)2)]．提案手法では，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 から成る英語コロケーションに翻訳することを想定し，「n^J を v^J」の v^J と同じ意味の v^J と強く共起するいくつかの日本語名詞 $N_1^J, \cdots, N_k^J$ を手掛かりに，その訳語 $N_1^E, \cdots, N_k^E$ と強く共起する英語動詞をweb文書中から抽出し，その中から v^J の妥当な訳語を求める．

提案手法は，辞書を用いただけでは妥当な訳語が判断できない場合におけるコロケーション翻訳に特に有効だと思われるが，その点に関する本手法の有用性についてはまだ検証が行なわれていない．そこで，本稿では実際に英語文書を作成するのに際し，辞書だけで対処できない場合がどのくらい起こりうるかを調査し，またそのときに本手法がどのくらい有効であるかを検証する．なお，実際の英語文書の作成を，日本語論文を英語に翻訳するというタスクに置き換え，提案手法の有用性を検証する．

本稿の構成は次の通りである．まず，2.章において提案手法の概要について説明し，アルゴリズムを示す．その後，3.章において実際の英語文書作成について本手法の有用性に対する調査，考察を行なう．最後に4.章でまとめを行なう．

2. 提案手法

2.1 提案手法の概要

まず，提案手法の概要について簡単に述べる．なお，提案手法の詳細については文献 2) を参照していただきたい．

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 から成る英語コロケーションに翻訳することを考える．ただし，n^J に対する妥当な訳語 n^E は予め分かっているものとする．また，「n^J を

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．

v^J」の共起性は比較的大きいものとする．提案手法では，このようなコロケーションに対して「n^J を v^J」の v^J に対する妥当な訳語を求めることを目的とする．

本手法の基本アイデアは次の通りである．まず，翻訳対象「n^J を v^J」に対して，日本語側の n^J と v^J とが強く共起するのならば，それを翻訳したときの英語側での n^E と v^E も強く共起するものと考えられる．また，ここで「n^J を v^J」の v^J と同じ意味の v^J と強く共起する日本語名詞の集合 Γ_J を考えると，「n^J を v^J」および各「N^J を v^J」($N^J \in \Gamma_J$) の v^J に対する妥当な訳語は，共通の動詞 v^E である可能性が高い．さらに，$N^J \in \Gamma_J$ と v^J とが強く共起するとき，多くの N^E (N^E は N^J の訳語）において，N^E と v^E も強く共起する傾向にある．つまり，「n^J を v^J」の v^J に対する訳語 v^E は n^E および N^E の多くと共通に強く共起する動詞の中に含まれていると考えられる．

ここで，単語 w が関係 f で単語 w' に係っているというコロケーションを $\langle w, f, w' \rangle$ と表記する．また，$\langle w, f, w' \rangle$ の共起性 $C(w, w' \mid f)$ は，文献 3) などで用いられている相互情報量

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

を基に評価する．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．

提案手法では，上記の基本アイデアにしたがって，n^J の訳語 n^E と強く共起する英語動詞をweb文書から抽出し，それらの動詞と$N^J \in \Gamma_J$ の各訳語とのweb上での共起性を計算して，Γ_Jの要素のうちでどのくらいの名詞と共通に強く共起するかを評価に，「n^J を v^J」に対する妥当な訳語候補を求め，優先度付でその訳語候補を提示する．

2.2 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．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 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．

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

   ここで $trans(N^J)$ はN^Jの訳語を表す．
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 $E(V^E)$

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

   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. 調　　査

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に対する翻訳に特に有効であると思われる．本章では，非母語話者が実際に英語文書を作成することを想定し，英語文書作成の際に，辞書だけで対処できない場面がどのくらい起こり得るかを調査する．また辞書で対処できない場合に本手法でどのくらい妥当な訳語を求めることができるかを検証することにより，本手法の有用性を示す．なお，今回は，英語文書の作成というタスクを，情報処理学会論文誌に掲載された日本語論文を英語に翻訳するというタスクに置き換え，調査を行なう．

3.1 辞書だけで妥当な訳語が得られる可能性に関する調査

日本人が英語文書を作成するときに，意図した日本語コロケーションを英語コロケーションに翻訳することを考える．日本語を英語に翻訳するには，和英辞書や英和辞書を用いて妥当な訳語を求めることが考えられるが，実際には辞書だけでは妥当な訳語を正確に判断できない場合も多い．そこで，妥当な訳語を求めるのに辞書だけでどのくらい対処可能であるかを次のように調査する．

この調査では，情報処理学会論文誌に掲載された日本語論文[3]を英語に翻訳することを想定し，その中の日本語名詞 n^J が格助詞「を」を介して日本語動詞 v^J に係るような日本語コロケーションを，英語名詞 n^E が目的語と

して英語動詞 v^E に係るような英語コロケーションに変換するというタスクを設定する．

まず，論文中で，名詞 n^J が格助詞「を」を介して動詞 v^J に係るような日本語コロケーションをすべて抜き出す．抜き出した各日本語コロケーション$\langle n^J, 『を』, v^J\rangle$における動詞 v^J の妥当な訳語を求めるために，次のように辞書を引く．

1. 動詞 v^J を和英辞書で引き，その中の例文に$\langle n^J, 『を』, v^J\rangle$（あるいは$\langle n^J, 『を』, v^J\rangle$と意味的にほぼ同等な日本語コロケーション）があれば，その英訳$\langle trans(n^J), \mathbf{obj}, V^E\rangle$から$V^E$を$v^J$の訳語候補として抜き出す．
2. 名詞 n^J を和英辞書で引き，その中の例文に$\langle n^J, 『を』, v^J\rangle$（あるいは$\langle n^J, 『を』, v^J\rangle$と意味的にほぼ同等な日本語コロケーション）があれば，その英訳$\langle trans(n^J), \mathbf{obj}, V^E\rangle$から$V^E$を$v^J$の訳語候補として抜き出す．
3. 名詞 n^J の訳語 $trans(n^J)$ を英和辞書で引き，その中の例文に$\langle n^J, 『を』, v^J\rangle$（あるいは$\langle n^J, 『を』, v^J\rangle$と意味的にほぼ同等な日本語コロケーション）があれば，その英訳$\langle trans(n^J), \mathbf{obj}, V^E\rangle$から$V^E$を$v^J$の訳語候補として抜き出す．

このうち，2，3で訳語候補が見つかった場合，その訳語候補は妥当なものだといえる．それに対し，1で見つけた訳語候補については，辞書の例文に翻訳対象の日本語コロケーション（あるいは翻訳対象と意味的にほぼ同等だと思われる日本語コロケーション）が使われていれば，妥当な訳語だと見なしてよいが，そうでない場合には妥当かどうかの判断は人間の内省に頼ることになる．

そこで，テストデータを

Case 1. 辞書の例文中で妥当な訳語がみつかったもの，

Case 2. 辞書の例文には載っていなかったが，人間の内省により（辞書の例文も参考にして）妥当な訳語が求まったもの，

Case 3. 人間の内省ではどれが妥当な訳語かを判断できなかったもの，あるいは妥当な訳語が見つからなかったもの，

に分類し，妥当な訳語を求めるのに辞書でどれだけ対処できるかを調べる．

翻訳対象とした日本語論文[4]について，そこから抽出した日本語コロケーション102個[†2]について調査を行ったところ，その内訳は**Table 1**のようになった． 表中での“Case i.”はそれぞれ上記の訳語候補の分別に対応する．**Table 1**を見ると，実際に英語で文書を書く際に，著者の意図する日本語コロケーションが辞書の例文にそのまま現れることはほとんどないと言ってよい．また，全体の7割程度のコロケーションについては，辞書だけを頼りにしても妥当な訳語を得ることができた．なお，今回の調査では論文を翻訳対象にしたため，特徴として「抽出する」，「構築する」のようなサ変動詞が多く見受けられた．サ変動詞に対する訳語は共起する名詞によって変化することが少なく，そのためv^J を和英辞書で調べるだけで対処できる場合が多く，Case 2. の割合が大きくなったものと思われる．これに関しては散文やニュースなどの文書形態によっても，割合が変わってくるのではないかと推測される．Case 2. の割合は予想よりも多かったものの，残りの3割については辞書だけを頼りにするだけでは妥当な訳語を判断することができず，本手法が有用となる場面は，多く残されていると言える．

Table 1 Examination result of seeking out the proper translation using dictionaries.

Case 1.	Case 2.	Case 3.	total
3	73	26	102

3.2 提案手法での訳語抽出に対する有効性の調査

前節で調査した日本語コロケーションに対し，それぞれのCaseについて提案手法を用いてweb上から訳語候補を抽出し，手法の有効性を検証する．

調査にあたり，提案手法を次のように適用する．まず，Γ_J について考える．提案手法で得られる訳語候補は，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「n^J を v^J」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，各「n^J を v^J」について，システムが提示した日本語名詞候補を，共起性の大きかったものから順に，「n^J を v^J」と同義な v^J として共起するかを調べながら，$|\Gamma_J|$ が10程度となるように名詞集合を用意することにする．このように準備したデータを用いて，提案手法で訳語候補を求める．

次に，閾値 θ_E について考える．本システムの有効性は正解率のみで評価はできない．閾値 θ_E を大きくとれば，システムが提示する候補の中に妥当な訳語が含まれる可能性は高くなるが，代わりに候補の絞込みを十分行なうことができず，利用者は多くの動詞について訳語としての妥当性をチェックする必要があり，利用者への負担が大きくなる．逆に閾値 θ_E を小さくとれば，チェックを要する訳語候補数は少なくなるが，その中に妥当な訳語が含まれる可能性も小さくなる．これを踏まえて，閾値

†2 論文中で現れた日本語コロケーションの異なり数である．論文中で複数回現れたコロケーションについては，同じコロケーションの訳語を複数回調べることはないとして1個として扱う．

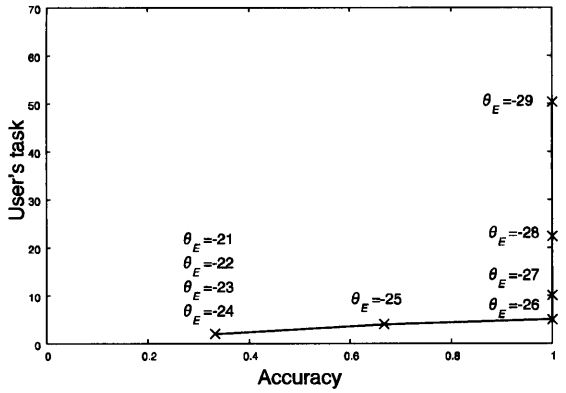

Fig. 1 Relation between accuracy and user's task in Case 1.

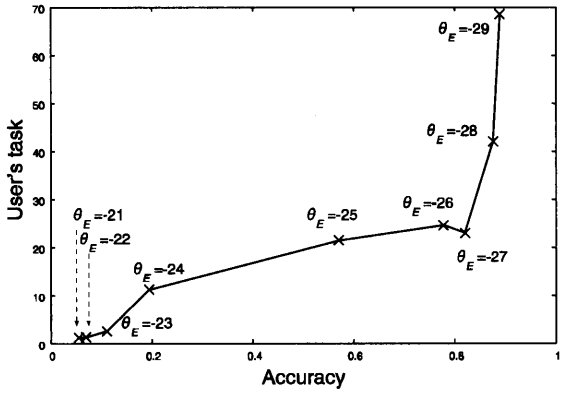

Fig. 2 Relation between accuracy and user's task in Case 2.

θ_E を適当な値に調整する必要があるが，それには以下の2点を考慮すべきである．

- 正解率が高い，つまりある程度まで調べれば妥当な訳語が見つかる可能性が高いこと．
- 正解が見つかるとして，そのとき利用者の負担が小さい，つまり妥当な訳語を見つけるのに調べなければならない訳語候補が少なくて済むこと．

そこで，閾値 θ_E を変化させながら，正解率と利用者の負担の程度を比較して，本手法の性能評価を行なうことにする．

正解率と利用者への負担の程度との関係を次のように調査する．まず，あるθ_Eにおける正解率をΔに妥当な訳語が含まれている割合，すなわちΔ内に少なくとも一つは妥当な訳語が含まれているかどうかで評価する．また，そのとき，利用者に課せられる負担の程度として，システムが提示した訳語候補をEの高いものから順に調べ，平均的に何個まで調べれば最初の正解が見つかるか†3 を求める．前述の3つのCaseに関して，θ_Eを変えたときの正解率と利用者が見なければならなかった訳語候補数の平均との関係を**Fig. 1**–**Fig. 3**に示す．また，テストセット全体に関して正解率と利用者が見なければならなかった訳語候補数の平均との関係を**Fig. 4**に示す．

Fig. 1-**Fig. 3**を見ると，$\theta_E \geq -28$で利用者の負担が急激に増加することが分かる．また，**Fig. 1**と**Fig. 3**を比べると，θ_Eが小さいときにはCase 3の方がよく，利用者の負担に対する正解率の下がり方が小さくなっている．逆にθ_Eが大きいときには，若干ではあるがCase 2の方が正解率が高くなっている．結果を見ると，実際のシステムでは θ_E を -27 前後に設定するのがよいと思われるが，θ_E が -27 付近において，Case 2.とCase 3.とでは利用者の負担を考えるとCase 3.の方が若干性能がよく現れたものの，正解率においてはそれほど変わりなかった．

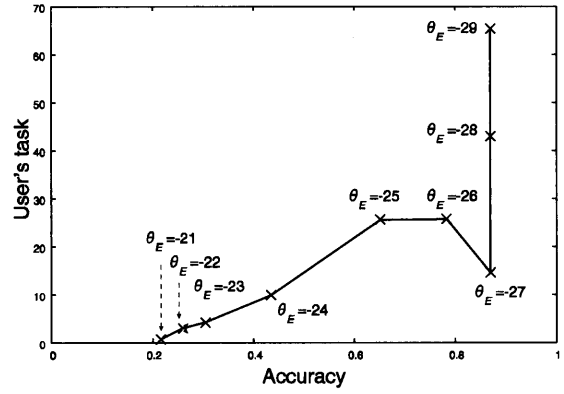

Fig. 3 Relation between accuracy and user's task in Case 3.

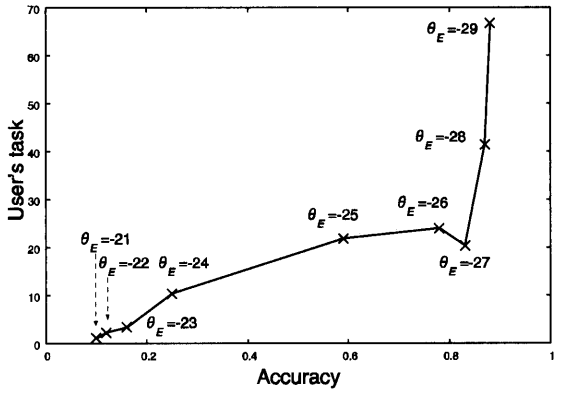

Fig. 4 Mean of the examination.

Fig. 4を見ると，$\theta_E = -27$のときに正解率0.83で，平均20.3個の訳語候補を調べればよいことになる．これらをすべて辞書や実文書で確認するのは，負担としては大きいと思われるが，候補には辞書を見なくても判断できるものも含まれており，実際の負担はこれよりも小さくなると推測される．

正解率と利用者の負担の程度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．実装に当たって適切な θ_E を求めるにはもっと大

†3 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する．

量のデータについて調査する必要があるものの，θ_E はシステム製作者が設定すべき値であり，θ_E の設定を利用者に求めるものではない．ただし，θ_E を変えたときに正解率と利用者の負担がどの程度になるのかを予め提示しておけば，それを見て利用者が θ_Eをカスタマイズすることは可能である．

4. おわりに

本稿では，web文書中の共起性に基づくコロケーション翻訳支援システムの有用性について調査を行なった．実際に英語文書を作成することを想定し，日本語論文中の「n^J を v^J」というタイプのコロケーションに対して翻訳を行なう場合について調査したところ，文書中の日本語コロケーションに対して，辞書だけでは妥当な訳語が判断できない場合が全体の約3割存在した．また，これらのコロケーションについて提案手法を適用して訳語候補を抽出したところ，辞書で対処できるものと辞書だけでは対処できないものとで，正解率においてはそれほどの差は見られなかった．利用者の負担においては，辞書で見つからないものの方が若干良い結果が現れた．全体としては$\theta_E = -27$のときに，正解率0.83，利用者の負担が20.3個となった．訳語候補のうちで，実際に辞書や実文書に当たって訳語の妥当性を確認する必要のあるものはこれよりも小さくなると思われるが，訳語候補の絞込みについては今後の検討課題である．

参考文献

1) 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**.
2) 柴田 雅博，冨浦 洋一，田中 省作: 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，**九州大学大学院システム情報科学紀要**, **10.1, 39–44, 2005**.
3) D. Hindle: Noun Classification from Predicate-Argument Structure, *Proc. 28th ACL*, **268–275, 1990**.
4) 南野 朋之，齋藤 豪，奥村 学: 繰り返し構造に基づいた Web ページの構造化，**情報処理学会論文誌**，**45.9, 2157–2167, 2004**.