隠れマルコフモデルに基づく日英専門用語対からの語構成要素対抽出 – 言語横断検索に向けて–

辻　慶太 * 影浦　峡 † 芳鐘冬樹 ‡

Abstract

言語横断検索ではキーワードの機械翻訳が一つの有効な方法である。本研究は，複合語キーワードの機械翻訳，翻訳用辞書の自動構築・維持に向けた基礎技術として，隠れマルコフモデルに基づき対訳複合語対から対訳語構成要素対を自動抽出する手法を提案する。対象言語は日本語・英語とし，人工知能用語 3742 対，ドキュメンテーション用語 1188 対を実験データとした。結果，日本語形態素解析ツールの援用とデータの前・後処理の併用により，延べの場合 90%以上の精度・再現率で対訳語構成要素対を抽出した。また訳語対抽出の先行研究でしばしば用いられる Dice 係数・相互情報量による抽出も行い，本モデルの有効性を比較確認した。

1　はじめに

世界規模で情報資源が蓄積される近年，言語の違いを越えた言語横断検索が注目を集めている。言語横断検索ではキーワードの機械翻訳が一つの有効なアプローチと考えられる。複合語キーワードの機械翻訳では，翻訳用辞書に複合語をそのままの形ですべて登録するのは現実的でなく，語構成要素を翻訳規則と共に登録するのが有効と思われる。そこで本研究では，複合語翻訳規則の抽出，あるいは複合語機械翻訳用辞書の自動構築・維持に向けた基礎技術として，隠れマルコフモデル (HMM) に基づき対訳用語対から，日本語語構成要素と英単語の対訳対 (以下「対訳単位対」) を自動抽出する手法を提案する。例えば，“並列マーカ伝播 (parallel marker propagation)”という対訳用語対から，‘並列 (parallel)’，‘マーカ (marker)’，‘伝播 (propagation)’ といった対訳単位対を自動抽出する手法である。このような研究はこれまでほとんど成されたことがない。

以下ではまず実験データを提示し，次に HMM に基づく対訳単位対抽出手法を説明する。さらに Dice 係数・相互情報量による抽出手法を取り上げ，結果の比較を行う。最後に，抽出に失敗する対の分析から，手法改良の可能性を検討したい。なお本研究では，英語における語構成要素とは単語を指すものとする。

2　データ

『人工知能大辞典』(丸善, 1991) の日英人工知能用語対 3742 個 (以下 AIT) と，『ユネスコ情報管理用語集』(丸善, 1984) の日英ドキュメンテーション用語対 1188 個 (以下 DCT) をデータとした。さて本研究手法は，日本語複合語の分割と英単語との対応付けを同時に行うのが特徴で，分割の為の日本語形態素解析ツールは本質的に必要としないが，それを援用した場合との結果比較も行う為，日本語部分については茶筌 1.51 で語構成要素分割を行った。

*東京大学大学院教育学研究科
E-mail:i34188@m-unix.cc.u-tokyo.ac.jp
†学術情報センター研究開発部
E-mail:kyo@rd.nacsis.ac.jp
‡東京大学大学院教育学研究科
E-mail:pp86037@hongo.ecc.u-tokyo.ac.jp

3　方法

本研究データには，日英の語構成要素同士が一対一に過不足無く対応する対訳用語対が多い。以下の HMM，Dice 係数・相互情報量による抽出では，この一対一対応を前提とする。その妥当性については後述する。

3.1　隠れマルコフモデルによる抽出

対訳用語対 “$J(E)$”で，日本語 J は m 文字 ‘$c_1 c_2 ... c_m$’，英語 E は n 単語 ‘$w_1 w_2 ... w_n$’ から成るとする。J を n 個の文字列に分割し，切り出された文字列を先頭から順に J_i とし，J_i における先頭から j 文字目を c_{ij} とする。この c_{ij} にはラベル Lj を与える。また J_1 ~ J_n のそれぞれに重複無く英単語を対応させ，J_i に w_k を対応させた場合，J_i の構成文字すべてに w_k を与える。このようにして出来たラベル・日本語文字・英単語の組み合わせを 1 つの状態と考え，“$J(E)$”内の語構成要素及び日英対応をこれら状態の遷移ととらえる。例えば “解析木 (analysis tree)”という対訳用語対には，図 1 のような状態遷移列が与えられる (ここで IF は始状態と終状態を表し，ϕ_J，ϕ_E はそれぞれ日本語の Null 文字，英語の Null 単語とする)。各状態間の遷移確率は後述のアルゴリズムで計算し，確率積が最大になる状態遷移列を，その対訳用語対の最適な分割・対応付けとする。例えば，表 1 で状態遷移列 (3) の確率が最大になったならば，“解析木 (analysis tree)”は ‘解析 (analysis)’，‘木 (tree)’ という対訳単位対

から成ると判断される。

(1)	IF	→	L1	→	IF	→	L1	→	L2	→	IF
	ϕ_J		解		ϕ_J		析		木		ϕ_J
	ϕ_E		analysis		ϕ_E		tree		tree		ϕ_E
(2)	IF	→	L1	→	IF	→	L1	→	L2	→	IF
	ϕ_J		解		ϕ_J		析		木		ϕ_J
	ϕ_E		tree		ϕ_E		analysis		analysis		ϕ_E
(3)	IF	→	L1	→	L2	→	IF	→	L1	→	IF
	ϕ_J		解		析		ϕ_J		木		ϕ_J
	ϕ_E		analysis		analysis		ϕ_E		tree		ϕ_E
(4)	IF	→	L1	→	L2	→	IF	→	L1	→	IF
	ϕ_J		解		析		ϕ_J		木		ϕ_J
	ϕ_E		tree		tree		ϕ_E		analysis		ϕ_E

表 1: “解析木 (analysis tree)”に関する状態遷移列

本抽出手法は，漢語複合語の語構成要素分割を目的とした武田・藤崎 (1985) の手法を拡張したものである。主な変更点は，武田・藤崎 (1985) がラベルを L2 までに制限していたのを，本モデルでは 3 文字以上の外来語語構成要素も扱えるよう L3 以上も用いたこと，及び対訳単位対抽出のため，日本語文字だけでなく英単語も状態に組み入れた点である[1] 。

ではアルゴリズムの説明に移る。まず以下のように記号を定義する。

D :データ中の対訳用語対の数
W_i :データ中の i 番目の対訳用語対 ($1 \leq i \leq D$)
N_i :W_i に関する状態遷移列の数
S_{ij} :W_i に関する j 番目の状態遷移列 ($1 \leq j \leq N_i$)
n_{ij} :S_{ij} に含まれる状態遷移の延べ数
t_{ijk} :S_{ij} のうちの k 番目の状態遷移 ($1 \leq k \leq n_{ij}$)
U :全データの状態遷移の異なり数
T_m :全データの異なり状態遷移のうち m 番目の遷移 ($1 \leq m \leq U$)
$P_r(X)$:状態遷移あるいは状態遷移列 X の確率
$tmp(T_m)$:T_m に関する一時変数

最適な分割・対応付けは次のように与えられる。

(1) すべての $P_r(T_m)$ に 1 を代入する
(2) すべての $tmp(T_m)$ に 0 を代入する
(3) W_i に関するすべての $P_r(S_{ij})(= \prod_{k=1}^{n_{ij}} P_r(t_{ijk}))$ を計算する
(4) W_i に関するすべての $P_r(S_{ij})$ を $\sum_{j=1}^{N_i} P_r(S_{ij})$ で割る
(5) W_i に関するすべての S_{ij} 中の t_{ijk} に関する $tmp(t_{ijk})$ に $P_r(S_{ij})$ を加える
(6) (3) から (5) までの操作を全 $W_i(1 \leq i \leq D)$ に対して行う
(7) すべての $tmp(T_m)$ を $\sum_{m=1}^{U} tmp(T_m)$ で割る
(8) $P_r(T_m)$ と $tmp(T_m)$ の差が 0 に収束したら (9) へ，そうでなければ $P_r(T_m)$ に $tmp(T_m)$ を代入して (2) に戻る
(9) 各 W_i の S_{ij} のうち，$P_r(S_{ij})$ が最大の S_{ij} を，その W_i の最適な分割・対応付けとする

本研究では，茶筌の分割結果を参照した場合の効果を調べる為，茶筌の分割箇所以外で日本語を切る状態遷移列 S_{ij} の $P_r(S_{ij})$ に，(3) の段階で 10^{-7} をかけ，相対的に低い確率を与える場合 (以下 H1) と，かけない場合，即ち茶筌の結果を参照しない場合 (以下 H0) の 2 通りを考え，結果の比較を行う。

[1] HMM では状態から状態への遷移によって実際の文字等が出力されると考えるモデルが一般的だが，ここでは武田・藤崎 (1985) にならって状態自体を日本語文字と英単語の組み合わせとした。両者の形式は異なるが実質的には等価である。

3.2 Dice 係数・相互情報量による抽出

対訳コーパスからの訳語対抽出では，対訳文における共起度の高いペアを対訳用語対とする方法がしばしば用いられる。今回は共起度の尺度として，Dice 係数(Smadja ら (1996)，高尾ら (1996a)，北村 & 松本 (1997))と相互情報量 (Gale & Church (1991)，van der Eijk (1993)，Fung (1995)，Smadja ら (1996)，北村 & 松本 (1997)) を取り上げ，以下の対訳単位対抽出を行った。

まず u_j，u_e をそれぞれ日英の語構成要素とし，f_{11} は u_j と u_e が「共起した対訳用語対数」，$f_{1.}$，$f_{.1}$ はそれぞれ u_j，u_e が「出現した対訳用語対数」，$f_{..}$ は全対訳用語対数とする。この時，Dice 係数 DC と相互情報量 MI はそれぞれ次のように定義される。

$$DC = \frac{f_{11} \times 2}{f_{1.} + f_{.1}} \qquad MI = log \frac{\frac{f_{11}}{f_{..}}}{\frac{f_{1.}}{f_{..}} \frac{f_{.1}}{f_{..}}}$$

対訳用語対では対訳単位対は過不足無く対応していると仮定し，すべての分割・対応付けを考える。例えば “解析木 (analysis tree)”には，表 2 のように 4 通りの分割・対応付けを考える。本研究では，上記の「出現した対訳用語対数」を，< 分割・対応付け中の単位対に現れた回数 / 単位対の総数 > の全データに渡る合計，と定義し「共起した対訳用語対数」についても同様に定義する。例えば “解析木 (analysis tree)”のみにおける ‘解’ の出現頻度は 2/8，‘解’ と ‘analysis’ の共起頻度は 1/8 とした。このようにして全対訳単位対の DC・MI を計算し，DC (あるいは MI) の和が最大になる分割・対応付けをその対訳用語対の最適な分割・対応付けとした。

また HMM と同様，茶筌参照の効果を調べる為，茶筌の分割箇所以外で日本語を切る分割・対応付け中の単位に対して，その対訳用語対での出現・共起頻度に 10^{-7} をかける場合 (DC，MI に関してそれぞれ D1，M1 とする) と，かけない場合 (同 D0，M0) の 2 通りを考えた。

(1)	解 (analysis)	析木 (tree)
(2)	解 (tree)	析木 (analysis)
(3)	解析 (analysis)	木 (tree)
(4)	解析 (tree)	木 (analysis)

表 2: “解析木 (analysis tree)”に関する分割・対応付け

4 結果

まず抽出した対訳単位対の適否に関する判断基準を述べる。ここでは特定の辞書等は拠り所とせず基本的に筆者らの主観で判断した。しばしば問題となる単位の末尾については次の基準を設けた。(1) 切り出した日本語単位の末尾に ‘的’，‘の’，‘な’ といった接辞・助詞がある場合，訳とされた英単語が名詞の場合は対訳として不適切，名詞以外の場合は適切とした。(2) 名詞以外の英単語に対して，末尾に接辞・助詞のない日本語単位が抽出された場合，元の対訳用語対でも添えられていない場合は対訳として適切，添えられている場合は不適切とした。これらの基準から例えば，“概念的学習 (concept learning)”からの ‘概念的 (concept)’，及び “概念的学習 (conceptual learning)”からの ‘概念 (conceptual)’ は不適切とした。

結果は表 5 のようになった。ここで，TP:対訳用語対の数，CTP:正しく分割・対応づけられた対訳用語対の数，MP:抽出すべき対訳単位対の数，EMP:適否に関わ

らず抽出された単位対の数，CMP:正しく抽出された対訳単位対の数，である。また各方法の上段は全対訳用語対に関する結果，下段は英単語数 2 以上の対訳用語対に関する結果である。表 5 から，H1 は英単語数 2 以上の対訳用語対 2804 個のうち，2387 個 (85.13%) を正しく分割・対応づけ，また英単語数 2 以上の対訳用語対に含まれる 2766 個の異なり対訳単位対のうち，2356 個を抽出したこと (再現率 85.18%) 等が分かる。

表 5 から，HMM，Dice 係数・相互情報量による抽出のいずれも，茶筌の分割結果を参考にした方が，参考にしないよりも結果が良いことが分かる[2] 。また対訳用語対の分割・対応付け正解率及び抽出した対訳単位対の精度・再現率は，AIT・DCT 共に H1，即ち茶筌の分割結果を参考にする HMM が最も高く，その値は他のいずれの手法よりも有意水準 0.05 で高いことが分かった。また H0 の正解率・精度・再現率は，D0・M0 よりも有意水準 0.05 で高く，茶筌の分割結果を参考にしない条件下でも HMM が有効であることが示された。

4.1 一対一対応の検証と改良

以下では対訳用語対の対応度を検証し，それに基づく簡単な前・後処理をデータに施して結果向上を図りたい。まず，対訳用語対における語構成要素の対応を調べた所，表 3 のようになった。'1-1' は日本語語構成要素と英単語が一対一に過不足なく対応した対の数とその割合 (%) を表す。また 'ER' は "画像解析 (analysis of image)"のように，対応する日本語語構成要素がない英単語を含む対を，'JR' は逆に日本語語構成要素が余る対を表し，'EJR' は日英双方で語構成要素が余る対を表す。表 3 から，本研究データの対訳用語対の 93%程度は日英語が過不足なく対応した対であり，一対一対応を前提とした本研究抽出モデルが一定の妥当性を持つことが分かる。

データ	全対	1-1	ER	JR	EJR
AIT	3742	3506 (93.69%)	113 (3.02%)	185 (4.94%)	62 (1.66%)
DCT	1188	1116 (93.94%)	36 (3.03%)	64 (5.39%)	28 (2.36%)

表 3: 一対一対応の対及び対応しない語構成要素を含む対

JR，ER で余った日英語の異なり数及び，延べ数 1，2 位の単位は表 4 の通りである。AIT・DCT 共に，日本語では 'の'，英語では 'of' が突出して余りやすい。従って，これらに焦点を当てた前・後処理をデータに施すことで，結果の向上が期待できる。

'の'，'of'，'の (of)' の延べ数は，AIT ではそれぞれ 262，91，71 で，DCT では 23，12，6 であった。表 4 の延べ数も勘案すると，'の' は 'of' 以外とも対訳単位対を構成する場合が多いが，'of' は 'の' 以外とはほとんど対訳単位対を構成しないことが分かる。同様に 'a'，'the' は日本語と対訳単位対を構成することはないことも分かった。そこで以下の前処理 (R) をデータに施し，H0，H1 を実行してみた (以下 R0，R1)。また後処理 (S) を H0，H1 の結果に施して，'の' が余る対に対処し (以下 S0，S1)，さらに (R)(S) 両方を実施した場合も調べた (以下 A0，A1)。

(R) 英語部分から 'a'，'the' を除く。日本語が 'の' を含まない対訳用語対では，英語部分から 'of' を除く

(S) 抽出した単位対に，'J(E)' という対の方が 'J の (E)' よりも多く存在した場合，後者の 'の' は除く

結果は表 6 のようになった。(R)(S) いずれも若干の結果向上をもたらし，両方実施する場合が最も結果が良いことが分かる。

データ	言語	異なり	延べ 1 位	延べ 2 位
AIT	日	97	の (79)	的 (5)
	英	91	of (17)	a，the (5)
DCT	日	47	の (12)	を (3)
	英	25	of (4)	and，aspect，generic，body-punched (2)

表 4: 対訳用語中に対訳単位がない語構成要素

5 おわりに

本研究の結果，HMM による対訳単位対抽出が有効であることが示された。今後は (1) 前・後処理の充実，(2) 語構成要素同士の一対一対応を仮定しない抽出モデルの構築，(3) 二言語辞書の援用，を図り，対訳単位対のより正確な抽出を目指す。その上で言語横断検索に向けた複合語翻訳規則の抽出あるいは複合語機械翻訳システムの構築を行いたい。

参考文献

[1] Fung, P. (1995)"A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora." The 33rd Annual Conference of the Association for Computational Linguistics, p.233-236.

[2] Gale, W. A. and Church, K. W. (1991)"Identifying Word Correspondences in Parallel Texts." *Proceedings of DARPA Speech and Natural Language Workshop*, p.152-157.

[3] Shapiro, S. C. and Eckroth, D. [編]; 大須賀節雄 [監訳] (1991)『人工知能大辞典』丸善. [Shapiro, S. C. and Eckroth, D. (eds.) (1987) *Encyclopedia of Artificial Intelligence*, New York: J.Wiley & Sons.]

[4] Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) "Translating Collocations for Bilingual Lexicons: A Statistical Approach." *Computational Linguistics*, 22(1), p.1-38.

[5] van der Eijk, P. (1993) "Automating the Acquisition of Bilingual Terminology." *Proceedings of Sixth Conference of the European Chapter of the Association for Computational Linguistics*, p.113-119.

[6] Wersig, G. and Neveling, U. [編]; 日本ドクメンテーション協会 [訳] (1984)『ユネスコ情報管理用語集』丸善. [Wersig, G. and Neveling, U. (eds.) (1976) *Terminology of Documentation*, Paris: Unesco.]

[7] 北村美穂子，松本裕治 (1997)"対訳コーパスを利用した対訳表現の自動抽出." 情報処理学会論文誌, 38(4), p.727-736.

[8] 武田浩一，藤崎哲之助 (1985)"統計的手法を用いた漢字複合語の短単位分割." 自然言語処理, 48(2), p.1-8.

[2] 茶筌の分割「誤り」を調べたところ，一対一に対応した対訳用語対のうち，英単語との対応という点で切るべき箇所で茶筌が切っていなかった対は，AIT に 158 対，DCT には 77 対あった。そのうち他の箇所で英単語数以上に切り，必然的に分割・対応付けが失敗する対訳用語対は，AIT に 56 対，DCT に 4 対あった。

データ	方法	対訳用語対			対訳単位対									
					延べ					異なり				
		TP	CTP	正解率	MP	EMP	CMP	精度	再現率	MP	EMP	CMP	精度	再現率
AIT	H0	3742	3209	85.76	7112	7246	6240	86.12	87.74	3436	3762	2829	75.20	82.33
		2804	2271	80.99	6174	6308	5302	84.05	85.88	2766	3101	2169	69.95	78.42
	H1	3742	3325	88.86	7112	7246	6496	89.65	91.34	3436	3709	3018	81.37	87.83
		2804	2387	85.13	6174	6308	5558	88.11	90.02	2766	3046	2356	77.35	85.18
	D0	3742	2893	77.31	7112	7246	5527	76.28	77.71	3436	3918	2420	61.77	70.43
		2804	1956	69.76	6174	6308	4590	72.76	74.34	2766	3258	1761	54.05	63.67
	D1	3742	3210	85.78	7112	7246	6237	86.08	87.70	3436	3801	2875	75.64	83.67
		2804	2273	81.06	6174	6308	5300	84.02	85.84	2766	3140	2215	70.54	80.08
	M0	3742	2084	55.69	7112	7246	3749	51.74	52.71	3436	4825	1916	39.71	55.76
		2804	1147	40.91	6174	6308	2812	44.58	45.55	2766	4149	1241	29.91	44.87
	M1	3742	3097	82.76	7112	7246	5998	82.78	84.34	3436	3977	2833	71.23	82.45
		2804	2160	77.03	6174	6308	5061	80.23	81.97	2766	3315	2172	65.52	78.52
DCT	H0	1188	1069	89.98	2012	2060	1851	89.85	92.00	1046	1131	938	82.94	89.67
		792	673	84.97	1616	1664	1455	87.44	90.04	758	841	648	77.05	85.49
	H1	1188	1100	92.59	2012	2060	1924	93.40	95.63	1046	1118	998	89.27	95.41
		792	704	88.89	1616	1664	1528	91.83	94.55	758	826	706	85.47	93.14
	D0	1188	988	83.17	2012	2060	1687	81.89	83.85	1046	1169	834	71.34	79.73
		792	592	74.75	1616	1664	1291	77.58	79.89	758	881	546	61.98	72.03
	D1	1188	1075	90.49	2012	2060	1873	90.92	93.09	1046	1131	954	84.35	91.20
		792	679	85.73	1616	1664	1477	88.76	91.40	758	843	666	79.00	87.86
	M0	1188	797	67.09	2012	2060	1288	62.52	64.02	1046	1346	694	51.56	66.35
		792	401	50.63	1616	1664	892	53.61	55.20	758	1051	399	37.96	52.64
	M1	1188	1049	88.30	2012	2060	1816	88.16	90.26	1046	1180	952	80.68	91.01
		792	653	82.45	1616	1664	1420	85.34	87.87	758	891	663	74.41	87.47

表 5: HMM , Dice 係数・相互情報量による分割・対応付けの結果

データ	方法	対訳用語対			対訳単位対									
					延べ					異なり				
		TP	CTP	正解率	MP	EMP	CMP	精度	再現率	MP	EMP	CMP	精度	再現率
AIT	R0	3742	3222	86.10	7112	7215	6260	86.76	88.02	3436	3723	2837	76.20	82.57
		2804	2285	81.49	6174	6277	5323	84.80	86.22	2766	3062	2177	71.10	78.71
	R1	3742	3341	89.28	7112	7215	6522	90.40	91.70	3436	3669	3028	82.53	88.13
		2804	2404	85.73	6174	6277	5585	88.98	90.46	2766	3006	2366	78.71	85.54
	S0	3742	3220	86.05	7112	7246	6251	86.27	87.89	3436	3749	2826	75.38	82.25
		2804	2283	81.42	6174	6308	5314	84.24	86.07	2766	3088	2166	70.14	78.31
	S1	3742	3345	89.39	7112	7246	6516	89.93	91.62	3436	3687	3015	81.77	87.75
		2804	2408	85.88	6174	6308	5579	88.44	90.36	2766	3024	2353	77.81	85.07
	A0	3742	3234	86.42	7112	7215	6271	86.92	88.17	3436	3708	2833	76.40	82.45
		2804	2297	81.92	6174	6277	5334	84.98	86.39	2766	3047	2173	71.32	78.56
	A1	3742	3361	89.82	7112	7215	6542	90.67	91.99	3436	3647	3025	82.94	88.04
		2804	2424	86.45	6174	6277	5605	89.29	90.78	2766	2984	2363	79.19	85.43
DCT	R0	1188	1071	90.15	2012	2055	1854	90.22	92.15	1046	1124	940	83.63	89.87
		792	675	85.23	1616	1659	1458	87.88	90.22	758	834	650	77.94	85.75
	R1	1188	1102	92.76	2012	2055	1927	93.77	95.78	1046	1111	1000	90.01	95.60
		792	706	89.14	1616	1659	1531	92.28	94.74	758	819	708	86.45	93.40
	S0	1188	1069	89.98	2012	2060	1851	89.85	92.00	1046	1131	938	82.94	89.67
		792	673	84.97	1616	1664	1455	87.44	90.04	758	841	648	77.05	85.49
	S1	1188	1100	92.59	2012	2060	1925	93.45	95.68	1046	1117	998	89.35	95.41
		792	704	88.89	1616	1664	1529	91.89	94.62	758	825	706	85.58	93.14
	A0	1188	1071	90.15	2012	2055	1854	90.22	92.15	1046	1124	940	83.63	89.87
		792	675	85.23	1616	1659	1458	87.88	90.22	758	834	650	77.94	85.75
	A1	1188	1102	92.76	2012	2055	1928	93.82	95.83	1046	1110	1000	90.09	95.60
		792	706	89.14	1616	1659	1532	92.34	94.80	758	818	708	86.55	93.40

表 6: データに前・後処理を行った場合の HMM の結果