日英報道記事からの訳語対獲得における言語横断情報検索の有効性の評価*

堀内 貴司　日野 浩平　浜本 武　中山 健明
豊橋技術科学大学 工学部 情報工学系
{takashi,hino,hamamo,takeaki}@cl.ics.tut.ac.jp

宇津呂 武仁
京都大学大学院 情報学研究科
utsuro@i.kyoto-u.ac.jp

1　はじめに

近年，WWW 上の日本国内の新聞社などのサイトにおいては，日本語だけでなく英語で書かれた報道記事も掲載しており，これらの英語記事においては，同一時期の日本語記事とほぼ同じ内容の報道が含まれている．これらの日本語および英語の報道記事のページにおいては，最新の情報が日々刻々と更新されており，分野特有の新出語（造語）や言い回しなどの翻訳知識を得るための情報源として，非常に有用である．本研究では，これらの報道記事のページから日本語および英語など，異なった言語で書かれた文書を収集し，多種多様な分野について，分野固有の固有名詞 (固有表現) や事象・言い回しなどの翻訳知識を自動または半自動で獲得する手法についての研究を行う．

ここで，訳語対応の推定においては，従来，コンパラブルコーパスから訳語対応を獲得する目的で開発された手法 (例えば，[Rapp95, Fung98]) を適用することが考えられる．これらの手法では，いずれも，言語を横断して，訳語対候補の周辺文脈の類似性を測定することにより，訳語対応の度合を推定する．しかし，従来手法の問題点として，コンパラブルコーパスに含まれる単語・連語について，二言語間での全ての組合せを訳語対応の候補とするために，コーパスの規模が大きくなると，訳語対応の候補全てを計算対象とすることが困難となる．この問題を回避する方法の一つとして，本論文では，コンパラブルコーパスから内容的に関連のある記事のみを収集し，そこから訳語対応を推定するというアプローチをとる．これにより，訳語対応の候補数が削減されるため，訳語対候補の周辺文脈の類似性を計算することが現実的に可能となる．

本研究における日英関連報道記事からの翻訳知識獲得の流れを図 1 に示す [堀内 02, Utsuro02]．まず，翻訳知識獲得のための情報源収集を目的として，同時期に日英二言語で書かれた WWW 上の新聞社やテレビ局のサイトから，報道内容がほぼ同一もしくは密接に関連した日本語記事および英語記事を検索する [浜本 03]．そして，取得された関連記事対に対し，内容的に対応

*Evaluating Effects of Cross-Language IR in Bilingual Lexicon Acquisition from Japanese and English News Articles

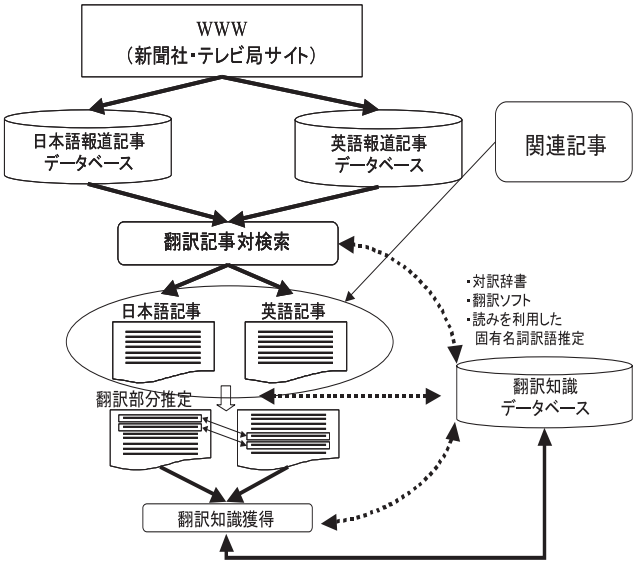

図 1: 日英関連報道記事からの翻訳知識獲得の流れ

する翻訳部分の推定を行い，その推定範囲から二言語間の訳語対応を推定し，訳語対の獲得を行う．特に，本論文では，この訳語対応の推定の過程において，言語横断情報検索によって関連記事対を絞り込むことの有効性を検証する．具体的には，関連記事対を絞り込むことにより，訳語対応推定の精度を落すことなく，訳語対応の候補を大幅に削減し，不要な計算が回避できることを示す．

2　日英関連報道記事における訳語対応の推定

本研究では，英語記事を検索質問として関連日本語記事を収集した結果から訳語対応を推定する [堀内 02, Utsuro02]．まず，検索質問となる英語記事を d_E^i として，d_E^i との間で余弦類似度の値が下限値 L_d 以上となる日本語記事の集合を D_J^i とする．

$$D_J^i \quad = \quad \left\{ d_J \mid \cos(d_E^i, d_J) \geq L_d \right\}$$

そして，D_J^i 中の記事を結合することにより一つの日本語記事 $D_J'^i$ を構成し，このような英日関連記事組 $\langle d_E^i, D_J'^i \rangle$ を集めた集合を RC_{EJ} とする．

$$RC_{EJ} \quad = \quad \left\{ \langle d_E^i, D_J'^i \rangle \mid D_J^i \neq \emptyset \right\}$$

本論文では，この関連記事組の集合 RC_{EJ} から訳語対応を推定する方法として，関連記事組の集合を疑似的

表 1: 単言語での単語・連語数および訳語対候補数

サイト			単言語における単語・連語の数		訳語対候補			対訳辞書中に存在する訳語対		
					訳語対数		割合 (full/ reduced)	訳語対数		割合 (full/ reduced)
			英語	日本語	reduced	full		reduced	full	
A	L_d (CLIR 利用)	0.5	780	737	52435	574860	11.0	141	285	2.0
		0.4	2684	3231	427889	8672004	20.3	543	1467	2.7
		0.3	5463	8119	1639714	44354097	27.1	1298	3492	2.7
	CLIR 利用せず		9265	65324	—	605226860	—	—	n/a	—
B	L_d (CLIR 利用)	0.5	2468	2158	494544	5325944	10.8	507	1206	2.4
		0.4	11968	8658	4074980	103618944	25.4	2155	n/a	—
	CLIR 利用せず		97998	71638	—	7020380724	—	—	n/a	—
C	L_d (CLIR 利用)	0.5	3760	2612	638089	9821120	15.4	753	1860	2.5
		0.4	13200	9433	4367775	124515600	28.5	2353	n/a	—
	CLIR 利用せず		119071	82055	—	9770370905	—	—	n/a	—

full: 英語単語または連語 t_E と日本語単語または連語 t_J のあらゆる組合せを訳語対候補とする場合
reduced: 擬似的対訳コーパス中の対訳文 (=関連記事の組) で観測された t_E と t_J の組合せのみを訳語対候補とする場合
n/a: 計算時間が大きいため計算を保留

な対訳コーパスとみなして，対訳コーパスにおける共起頻度を用いた訳語対応推定尺度を適用する方法，および，関連記事組の集合をコンパラブルコーパスとみなして，コンパラブルコーパスからの訳語対応推定手法を適用する方法の二種類を比較する．

以下，訳語対応推定の対象となる英語連語または単語を t_E，日本語連語または単語を t_J として，t_E と t_J の間の訳語対応推定値を $corr_{EJ}(t_E, t_J)$ とする．本論文では，t_E の品詞列としては任意のものを，また，t_J の品詞列としては，日本語形態素解析システム「茶筌」により品詞列を推定し，接頭詞，名詞，動詞によって構成される任意の列を対象としている[1]．さらに，t_E あるいは t_J が出現する記事数 $df(t_E)$ および $df(t_J)$ に下限 L_f^E および L_f^J を設け，また，英語連語および日本語連語を構成する単語数 $length(t_E)$ および $length(t_J)$ に上限 U_l^E および U_l^J を設ける．

$$df(t_E) \geq L_f^E,\ df(t_J) \geq L_f^J,\\ length(t_E) \leq U_l^E,\ length(t_J) \leq U_l^J \quad (1)$$

2.1 関連記事組における訳語候補の共起および分割表を用いた推定

関連記事組の集合 RC_{EJ} を疑似的な対訳コーパスとみなして訳語対応の推定を行う場合は，関連記事組の集合 RC_{EJ} 中の関連記事組 $\langle d_E^i, D_J'^i \rangle$ において t_E と t_J が共起する記事組数 $df(t_E, t_J)$ に下限 L^{EJ_f} を設け，2×2 分割表に基づく ϕ^2 統計値を $corr_{EJ}(t_E, t_J)$ とする．

$$df(t_E, t_J) \geq L^{EJ_f} \quad (2)$$

2.2 文脈の類似性を用いた推定

関連記事組の集合 RC_{EJ} をコンパラブルコーパスとみなして訳語対応の推定を行う場合は，t_E および t_J の各々について，それぞれが出現する文の頻度ベクトルを加算することにより，t_E および t_J に対する文単位の文脈頻度ベクトルし，文脈頻度ベクトル間の余弦を $corr_{EJ}(t_E, t_J)$ とする．ここで，訳語対応推定の対象となる t_E と t_J の組合せとしては，以下の二通りの方法を比較し，関連記事組の集合 RC_{EJ} における記事間の対応を利用することにより，訳語対応の候補を絞り込むことの有効性を評価する．

1. 式 (1) の制約を全ての t_E と t_J の組合せを訳語対応の候補とする (以下，“full”)．
2. 式 (1) および (2) の制約を満たす t_E と t_J の組合せを訳語対応の候補とする (以下，“reduced”)．

3 実験および評価

A∼C の三種類のサイトから日本語および英語の報道記事を収集し，記事間類似度の下限値 L_d のいくつかの設定のもとで，英語記事 100∼500 記事程度を検索質問とした言語横断関連報道記事検索により関連日本語記事を自動収集した (記事数の詳細は [堀内 02] を参照)．以下では，記事数下限値および連語を構成する単語数の上限値について，$L_f^E = L_f^J = 3, L_f^{EJ} = 2, U_l^E = U_l^J = 5$ という条件で，訳語対応の推定を行った．

3.1 訳語対候補数の比較

訳語対として推定される英語・日本語それぞれの訳語候補数と，既存の対訳辞書に存在する訳語数を表 1 に示す．「CLIR 利用せず」の行は，関連記事組の検索を行わず，英語/日本語報道記事コーパス全体を対象とした場合の統計値を示す．一方，「L_d(CLIR 利用)」の行は，言語横断関連報道記事検索を用い，記事検索の類似度閾値の下限を満たす関連記事組を絞り込んだ場合の統計値を示す．「CLIR 利用」と「CLIR 利用せず」における訳語対候補数を比較すると，言語横断関連報道記事検索を行わない場合の訳語対候補数のほうが圧倒的に大きいことがわかる．例えば，記事間類似度下限 $L_d = 0.5$ の場合との比較では，「CLIR 利用せず」の場合に約 1000 倍の訳語対候補が存在する．この結果より，本論文で対象とした WWW 上の日英報道記事

[1] 「茶筌」の品詞体系では，接尾辞は名詞に含まれる．

(a) サイト A, $L_d = 0.4$　(b) サイト B, $L_d = 0.5$　(c) サイト C, $L_d = 0.5$

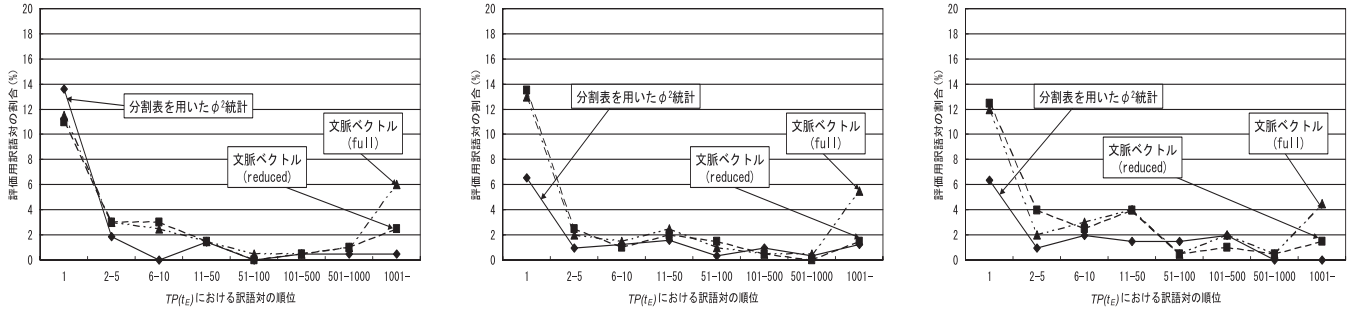

図 2: 評価用訳語対の割合 ($TP(t_E)$ 上位 200 個, $TP(t_E)$ における訳語対の順位ごと)

表 2: 訳語対応推定結果の人手による評価 (サイト A, 記事間類似度下限 $L_d = 0.4$, $TP(t_E)$,$TP_c(t_E)$ 上位 200 個)

訳語対応推定手法		評価用訳語対を含む $TP(t_E)$ の割合 (%) 関連記事組での訳語	関連記事組での訳語でない	人手により半自動獲得された訳語対数 (インデックス語: 最長語)
ϕ^2 統計		16.4 (35/213)	1.9 (4/213)	164 (内, 53 組=32.3%が対訳辞書に含まれず)
文脈ベクトル	reduced	21.5 (43/200)	6.0 (12/200)	116 (内, 63 組=54.3%が対訳辞書に含まれず)
	full	22.0 (44/200)	11.0 (22/200)	117 (内, 57 組=48.7%が対訳辞書に含まれず)

から訳語対を獲得するタスクに対して，コンパラブルコーパスから訳語対を獲得するための従来手法をそのまま適用すると，相対的な計算コストが大幅に増加することが分かる[2]．また,「訳語対候補数」の欄と「対訳辞書辞書中に存在する訳語対」の欄において「割合 (full/reduced)」の項目を比較すると，関連記事組を絞り込むことにより，不要な訳語対候補を大幅に除去できていることが分かる．これより，訳語対応の推定において，関連記事組を絞り込むことが有効であることが予測される．

3.2 訳語対応推定の性能

次に，訳語対応推定の性能を評価するために，本論文では，まず，推定された訳語対候補全体の集合を，英語連語または単語 t_E を共有する訳語対候補から構成される部分集合 $TP(t_E)$ に分割する．

$$TP(t_E) \quad = \quad \left\{ \langle t_E, t_J \rangle \mid t_E \text{を共有する訳語対候補 } \langle t_E, t_J \rangle \right\}$$

次に，各集合 $TP(t_E)$ に対して，要素となっている訳語組の訳語対応推定値 $corr_{EJ}(t_E, t_J)$ のうちの最大値を $corr_{EJ}(TP(t_E))$ とする．

$$corr_{EJ}(TP(t_E)) \quad = \quad \max_{\langle t_E, t_J \rangle \in TP(t_E)} corr_{EJ}(t_E, t_J)$$

そして，全ての集合 $TP(t_E^1), \ldots, TP(t_E^m)$ を $corr_{EJ}(TP(t_E))$ の値の降順に順位付けした上で，訳語対応推定の性能を評価する．

3.2.1 既存の対訳辞書中の訳語対応を用いた評価

既存の対訳辞書 (英辞郎 Ver.37: 85 万語) 中に含まれる訳語対を評価用訳語対として，上位 200 個の $TP(t_E)$ について，$TP(t_E)$ における訳語対の順位ごとに以下の割合を算出し，各順位における訳語対が既存の対訳辞書に含まれる「正しい」訳語である率を評価した．

$$\frac{|\{TP(t_E) \mid \text{評価用訳語対 } \langle t_E, t_J \rangle \in TP(t_E)\}|}{|\{TP(t_E) \mid \text{上位 200 位以内の } TP(t_E)\}|}$$

この結果をプロットしたものを図 2 に示す．訳語対応推定手法の間で性能を比較すると，サイト A では，$TP(t_E)$ 中の順位の上位での性能に大きな差はないものの，サイト B および C では，$TP(t_E)$ 中の順位の一位において，文脈ベクトルを用いる手法の性能が高い．また，文脈ベクトル (reduced) と文脈ベクトル (full) の性能を比較すると，いずれのサイトにおいても，$TP(t_E)$ 中の順位の下位における差が大きい．

そこで，次に，$TP(t_E)$ のそれぞれの順位に位置する評価用訳語対の特性を分析するために，サイト A, 記事間類似度下限値 $L_d = 0.4$ の場合について，各々の評価用訳語対が，関連記事組において実際に翻訳関係にあるか否かを人手で調査した．まず，上位 200 個の $TP(t_E)$ 全体での内訳を表 2 「評価用訳語対を含む $TP(t_E)$ の割合 (%)」の欄に示す．さらに，$TP(t_E)$ における訳語対の順位ごとにこの割合をプロットした結果を図 3 に示す．この結果から，文脈ベクトルを用いた訳語対応推定 (特に, "full" の場合) では，実際には関連記事組において翻訳関係になく，偶然に評価用訳語対と照合した訳語組が一定数含まれ，しかもそのほとんどが $TP(t_E)$ 中の下位に位置することが分かる．

以上の結果から，いずれの訳語対応推定手法を用いても，訳語対応推定値の上位に獲得の対象とすべき訳語対が位置しており，訳語対応推定値の信頼性が高いことが分かる．また，文脈ベクトル (reduced) と文脈

[2] 訳語対応推定の計算時間の見積もりでは，PentiumIV 1.9GHz の計算機を利用した場合，サイト A,「CLIR 利用せず」では約 6 日間，サイト B および C，記事間類似度下限値 $L_d = 0.4$ の場合の「CLIR 利用：full」でも約 6 日間，また，サイト B および C における「CLIR 利用せず」の場合は，半年以上を要する．

(a) ϕ^2 統計　　(b) 文脈ベクトル: reduced　　(c) 文脈ベクトル: full

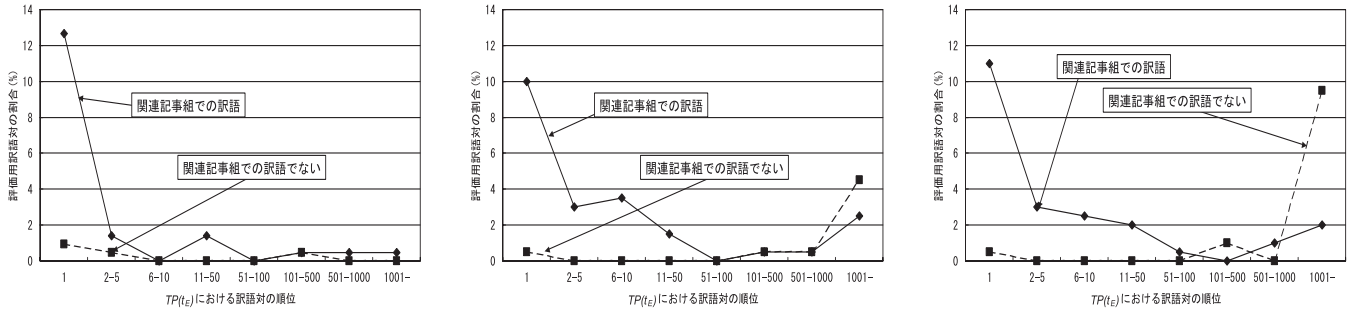

図 3: $TP(t_E)$ 中の評価用訳語対: 関連記事組での出現の有無の比較 (サイト A，記事間類似度下限 $L_d = 0.4$, $TP(t_E)$ 上位 200 個)

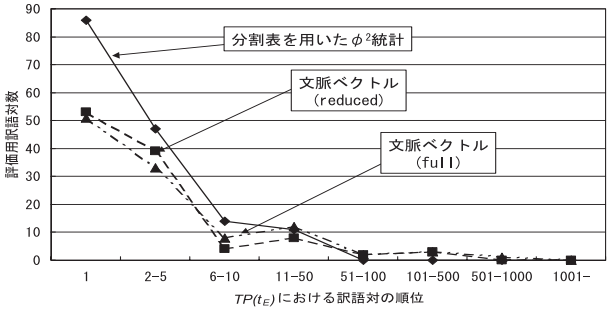

図 4: 人手により半自動獲得された訳語対数 (サイト A，記事間類似度下限 $L_d = 0.4$，インデックス語: 最長語，$TP_c(t_E)$ 上位 200 個)

ベクトル (full) の性能の間には大きな差がなく，計算量の大きい文脈ベクトル (full) を適用する必要性はないと言える．したがって，関連記事組を絞り込むことにより，訳語対応推定の性能を落すことなく不要な計算が回避できていると言える．

3.2.2 訳語対応推定結果の人手評価

次に，既存の対訳辞書の訳語対に限定せず，任意の訳語対応の半自動獲得 [日野 03] を行い，獲得された訳語対応と訳語対応推定値の相関について評価を行った．

まず，訳語対応半自動獲得の効率を上げるために，英語の連語もしくは単語 t_E に対して，t_E 自身もしくはその一部を構成する語を t'_E として，訳語組候補の集合 $TP(t'_E)$ の和集合を求め，これを $TP_c(t_E)$ とする [日野 03]．そして，サイト A，記事間類似度下限値 $L_d = 0.4$ の場合について，訳語対応推定値の最大値の降順に順位付けした上位 200 個の $TP_c(t_E)$ に対して，訳語対候補を人手で評価し，正しい訳語対であると判定された組を半自動獲得結果の訳語対とした．この訳語対数を表 2「人手により半自動獲得された訳語対数」の欄に示す．また，$TP(t_E)$ における訳語対の順位ごとに，獲得された訳語対数をプロットした結果を図 4 に示す．この結果から，特に，ϕ^2 統計を用いて訳語対応推定を行った場合に，既存の対訳辞書に含まれる訳語対が多く獲得されているが，全体として，どの訳語対応推定手法を用いても，既存の対訳辞書に含まれない新規の訳語対応がほぼ同程度獲得できている．また，文脈ベクトル (reduced) と文脈ベクトル (full) の性能の間には大きな差が見られない．実際，獲得された訳語対の重複を分析したところ，文脈ベクトル (reduced) と文脈ベクトル (full) の間では，ほぼ全ての訳語対が重複しており，文脈ベクトル (full) を適用する必要がないことが分かった．一方，ϕ^2 統計と文脈ベクトル (reduced) との間で獲得された訳語対の重複を分析したところ，ϕ^2 統計において獲得された訳語対の約 72%が重複しておらず，両者は相補的な特性を持つことが分かった．

4 おわりに

本論文では，日英関連報道記事からの訳語対応推定において，言語横断情報検索によって関連記事対を絞り込むことの有効性を検証した．特に，関連記事対を絞り込むことにより，関連記事対を絞り込むことにより，訳語対応推定の精度を落すことなく，訳語対応の候補を大幅に削減し，不要な計算が回避できることを示した．

参考文献

[Fung98] Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 17th COLING and 36th ACL*, pp. 414–420 (1998).

[浜本 03] 浜本武, 中山健明, 日野浩平, 堀内貴司, 宇津呂武仁：言語横断関連報道記事検索における翻訳ソフト・対訳辞書・数値表現翻訳規則の性能比較, 言語処理学会第 9 回年次大会論文集 (2003).

[日野 03] 日野浩平, 堀内貴司, 浜本武, 中山健明, 宇津呂武仁：日英関連報道記事からの翻訳知識獲得のためのユーザインタフェースの作成, 言語処理学会第 9 回年次大会論文集 (2003).

[堀内 02] 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).

[Rapp95] Rapp, R.: Identifying Word Translations in Non-Parallel Texts, *Proc. 33rd ACL*, pp. 320–322 (1995).

[Utsuro02] Utsuro, T., et al.: Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites, *Machine Translation: From Research to Real Users*, Lecture Notes in Artificial Intelligence: Vol. 2499, pp. 165–176, Springer (2002).