2004－NL－162　(9)
2004／7／15

日英報道記事からの訳語対応推定: ターム頻度と訳語対応推定性能の相関の評価

日野　浩平† 　宇津呂武仁‡ 　中川　聖一†

† 豊橋技術科学大学 工学部 情報工学系
‡ 京都大学大学院 情報学研究科 知能情報学専攻
†{hino, nakagawa}@slp.ics.tut.ac.jp, ‡utsuro@i.kyoto-u.ac.jp

近年，ウェブ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．本研究では，これらの報道記事のページから，日本語で書かれた文書および英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というアプローチをとる．翻訳知識獲得においては，まず，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．そして，関連記事組における訳語候補の共起に基づく相関尺度を用いて，二言語間の訳語対応を推定する．本稿では，この尺度を用い，英語タームの出現頻度の分布に応じて，訳語対応推定性能が変化するかどうかを調査し，その相関を評価する．そして，英語タームの頻度が大きいほど，高い訳語対応推定性能が達成できることを示す．

Estimating Bilingual Term Correspondences from Japanese-English News Articles:

Evaluation on Correlation of Term Frequencies and Correctness of Term Correspondences Estimation

Kohei HINO†, Takehito UTSURO‡, and Seiichi NAKAGAWA†

† Dpt. Information and Computer Sciences, Toyohashi University of Technology
‡ Dpt. Intelligence Sci. and Tech., Graduate School of Informatics, Kyoto University
†{hino, nakagawa}@slp.ics.tut.ac.jp, ‡utsuro@i.kyoto-u.ac.jp

This paper focuses on bilingual news articles on WWW news sites as a source for translation knowledge acquisition. We take an approach of acquiring translation knowledge of domain specific named entities, event expressions, and collocational expressions from the collection of bilingual news articles on WWW news sites. In this framework, pairs of Japanese and English news articles which report identical contents or at least closely related contents are retrieved. Then, a statistical measure is employed for the task of estimating bilingual term correspondences based on co-occurrence of Japanese and English terms across relevant Japanese and English news articles. This paper then examines the correlation of term frequencies and correctness of term correspondences estimation. We experimentally show that the more frequent the target English terms be, the more reliably bilingual term correspondences can be estimated.

1. はじめに

近年，ウェブ上の日本国内の新聞社などのサイトには，日本語だけでなく英語で書かれた報道記事も掲載されており，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) や新出複合語，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．そこで，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というアプローチをとる [9].

これまでに研究されてきた翻訳知識獲得の手法は，大きく，対訳コーパスからの獲得手法とコンパラブルコーパスからの獲得手法に分けられる [5]．通常，対訳コーパスにおいては，文の対応の情報を利用することにより，片方の言語におけるタームや表現について，もう一方の言語における訳の候補が比較的少数に絞られるため，翻訳知識

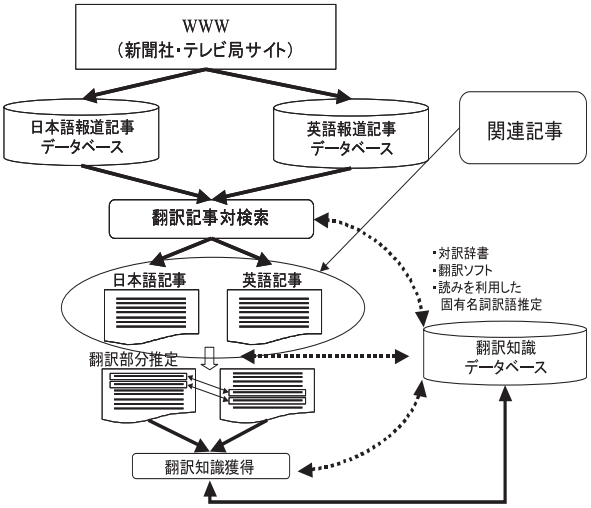


図 1　日英関連報道記事からの翻訳知識獲得の流れ

の獲得は相対的には容易といえる．ただし，そのような対訳コーパスを人手で整備する必要がある点が短所である．一方，コンパラブルコーパスからの獲得 (例えば, [1], [7]) では，各タームの周囲の文脈の類似性を言語横断して測定することにより，訳語対応の推定が行われる．情報源となるコーパスを用意するコストは小さくて済むが，翻訳知識の獲得は相対的に難しく，高性能に翻訳知識獲得を行うのは容易ではない．これらの従来のアプローチと比べると，報道記事を用いる本研究のアプローチは，情報源となるコーパスを用意するコストについては，コンパラブルコーパスを用いるアプローチと同等に小さく，しかも同時期の報道記事を用いるため，片方の言語におけるタームや表現の訳がもう一方の言語の記事の方に出現する可能性が高く，翻訳知識の獲得が相対的に容易になるという大きな利点がある．

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 [9]．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ウェブ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 [2]．そして，取得された関連記事対に対し，内容的に対応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．ここで，訳語対応を推定する尺度としては，関連記事組における訳語候補の共起を利用する方法の有効性をすでに示してきた [9]．本稿では，この尺度を用い，英語タームの出現頻度の分布に応じて，訳語対応推定性能が変化するかどうかを調査し，その相関を評価した．その結果，英語タームの頻度が大きいほど，高い訳語対応推定性能が達成できることが分かった．

2. 言語横断関連報道記事検索

言語横断関連報道記事検索においては，まず，新聞社やテレビ局のサイトから英語記事 d_E と日本語記事 d_J を取得する．次に，関連記事対はお互いの日付が近いと想定して，日付の情報を用いて検索対象の記事を絞りこむ．そして，取得した英語記事 d_E と日本語記事 d_J の間の類似性を測るために，翻訳ソフト・対訳辞書などの情報源(注1)を利用して英語記事 d_E を日本語訳に変換し，この日本語訳と日本語記事 d_J から翻訳頻度ベクトル $v_{trJ}(d_E)$ と日本語頻度ベクトル $v(d_J)$ をそれぞれ作成する(注2)．最後に，頻度ベクトル間で類似度を計算し，類似度が下限値以上の記事を検索結果とする．

表 1　2×2 分割表

	t_J	$\neg t_J$
t_E	$df(t_E, t_J) = a$	$df(t_E, \neg t_J) = b$
$\neg t_E$	$df(\neg t_E, t_J) = c$	$df(\neg t_E, \neg t_J) = d$

ここで，この検索結果から，日英関連記事組を作成する場合には，英語記事を検索質問として関連日本語記事を収集する場合と，逆に，日本語記事を検索質問として関連英語記事を収集する場合の二通りが考えられる．英語記事を検索質問として関連日本語記事を収集する場合は，検索質問となる英語記事 d_E の日本語訳頻度ベクトル $v_{trJ}(d_E)$ 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 とする．

$$D_J = \left\{ d_J \mid \cos(v_{trJ}(d_E), v(d_J)) \geqq L_d \right\}$$

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

$$RC_{EJ} = \left\{ \langle d_E, D'_J \rangle \mid D_J \neq \emptyset \right\}$$

3. 日英関連報道記事における訳語対応の推定

本稿では，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．以下，訳語対応推定の対象となる英語ターム (連語または単語) を t_E，日本語ターム (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．本稿では，t_E の品詞列

(注1)：翻訳ソフト (オムロン社製「翻訳魂」) と対訳辞書 (英辞郎 Ver.37, 85 万語) を比較した結果では，翻訳ソフトの方が高い検索性能を達成しており，そのため，訳語対応推定においても，翻訳ソフトを用いた方が高い性能が得られている．そこで，本稿では，翻訳ソフトを用いて英語記事の日本語訳を行った後，関連記事検索を行った結果を用いる．

(注2)：日本語形態素解析システム「茶筌」(http://chasen.aist-nara.ac.jp/) を用いて形態素列に分割し，平仮名語の高頻度機能的表現 26 語を不要語として削除した．また，単語頻度ベクトルは名詞と動詞のみを利用して生成した．なお，5 形態素以下の連語および単語の頻度を次元として頻度ベクトルを構成する場合と，単語の頻度のみを次元として頻度ベクトルを構成する場合を比較すると，翻訳ソフトを用いた場合では大きな差はない．しかし，対訳辞書を用いた場合は，連語および単語の頻度を次元とする方が性能がよいため，こちらを用いることとする．

表 4　評価用英語ターム数の分布

サイト	全体					ϕ^2統計値 1∼0.15			ϕ^2統計値 0.15∼0.07			ϕ^2統計値 上位 100		
	頻度	5∼10	10∼20	20 以上	頻度	5∼10	10∼20	20 以上	5∼10	10∼20	20 以上	5∼10	10∼20	20 以上
A	MT	117	158	531	MT	58	82	289	32	37	147	38	110	103
	辞書	1391	1718	2507	辞書	285	407	727	229	304	570	73	166	157
					人手	148	116	131	51	48	56	100	100	100
	その他	4423	3483	2786	除外	866	671	687	800	684	618	199	75	66
	総数	5931	5359	5824	総数	1357	1276	1834	1112	1073	1391	397	381	360
B	MT	104	214	791	MT	87	124	377	28	57	226	87	128	95
	辞書	1098	1367	1968	辞書	216	321	590	203	236	452	216	333	218
					人手	104	71	102	25	45	26	100	100	100
	その他	3105	2364	1868	除外	669	476	462	570	432	418	673	487	306
	総数	4292	3835	4167	総数	1048	922	1298	808	740	995	1048	977	668
C	MT	103	164		MT	75	114		22	43		103	164	
	辞書	226	205		辞書	147	125		46	60		226	205	
					人手	43	35		10	4		57	40	
	その他	313	152		除外	158	68		54	33		256	112	
	総数	585	424		総数	379	275		123	115		585	424	

表 2　記事の日数・記事数・平均記事長

新聞社		総日数	総記事数	一日の平均記事数	一記事の平均記事長 (byte)
サイト A	英語	935	23064	24.7	3228.9
	日本語	941	96688	102.8	837.7
サイト B	英語	935	14587	15.6	3302.6
	日本語	941	81652	86.8	867.9
サイト C	英語	935	1553	1.6	1368.6
	日本語	941	9660	10.2	774.3

表 3　記事間類似度の下限を満たす日英報道記事の数

	類似度下限 L_d	CLIR	0.3	0.4	0.5
サイト A	日付幅 (日)	なし	± 2		
	英語記事数	23064	6073	2392	701
	日本語記事数	96688	12367	3444	882
	日本語記事数 (重複あり)		16507	3840	918
	類似度下限 L_d	CLIR	0.3	0.4	0.5
サイト B	日付幅 (日)	なし	± 2		
	英語記事数	14587	4316	1658	396
	日本語記事数	81652	8108	2349	499
	日本語記事数 (重複あり)		11451	2694	523
	類似度下限 L_d	CLIR	0.3	0.4	0.5
サイト C	日付幅 (日)	なし	± 4		
	英語記事数	1553	765	413	159
	日本語記事数	9660	1918	673	192
	日本語記事数 (重複あり)		2406	766	203

としては任意のものを，また，t_J の品詞列としては,「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．

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際には，RC_{EJ} 中の関連記事組 $\langle d_E, D'_J \rangle$ において t_E と t_J が共起する記事組数 $df(t_E, t_J)$，t_E のみが含まれ t_J が含まれない記事組数 $df(t_E, \neg t_J)$，t_J のみが含まれ t_E が含まれない記事組数 $df(\neg t_E, t_J)$，t_E も t_J も含まれない記事組数 $df(\neg t_E, \neg t_J)$ を用いて表 1 の 2×2 分割表を構成する．一般に共起推定でよく用いられる相互情報量，ϕ^2 統計，dice 係数，対数尤度比などの尺度を比較したところ，訳語対応推定の性能としては，ϕ^2 統計，が最もよく，対数尤度比,dice 係数，相互情報量という順で性能が劣化した [3]．そこで，以下の ϕ^2 統計を用いて t_E と t_J の統計的相関を測定し，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ とする．

$$\phi^2(t_E, t_J) = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

4. 実験および評価

4.1 報道記事セット

国内の新聞社等 (今回は三社) のウェブサイトから，表 2 に示す日数・記事数・記事長の英語および日本語の報道記事を収集した．次に, [3] における調査結果に基づいて，英語の記事に対してほぼ同一の内容の日本語記事が存在する日付の幅を設定し，その日付の幅の範囲で言語横断関連報道記事検索を行った．記事間類似度下限 L_d を変化させた場合に検索される記事数の一覧を表 3 に示す．ここで,「日本語記事数 (重複あり)」の欄には，二つ以上の英語記事に対して重複して検索された日本語記事を重複して数えた記事数を示す．この結果から，類似度下限 L_d が 0.4 や 0.5 の場合は，利用可能な記事数が著しく減少することが分かる．予備実験において，訳語対応推定が安定して行えるためには，一定規模以上の記事が必要であるという結果が得られていたため，以降の訳語対応推定は，類似度下限 $L_d = 0.3$ の条件のもとで行う．

4.2 ターム頻度分布ごとの評価用英語タームの選定

今回の評価実験では，評価用英語ターム (機械翻訳で訳せず，対訳辞書辞書に存在しない単語または連語) を人手で選定しておき，評価用英語タームに対する日本語訳語候補の順位付けの性能の評価を行った．実装の都合上，英語タームおよび日本語タームを構成する単語数に

表 5　評価用日英ターム組の例

英語ターム	日本語ターム
High Public Prosecutors Office	高検
Environment Ministry	環境省
Japanese Consulate General	日本総領事館
diesel-powered vehicles	ディーゼル車
Japan Coast Guard	海上保安庁
fertilized eggs	受精卵
Tokyo District Public Prosecutors Office	東京地検
Aum Supreme Truth	オウム真理教
intellectual property rights	知的財産権
special structural reform zones	構造改革特区

上限 U_l^E および U_l^J を設け，$U_l^E = U_l^J = 5$ とした．ここで，英語ターム出現頻度の計算を効率よく行うために，PrefixSpan [6](注3) を用いて頻度 5 以上の単語列の頻度を測定した．この頻度分布で, 頻度 5 以上 10 未満，10 以上 20 未満，20 以上，の三種類の出現頻度分布で単語列集合を分割し, それぞれの集合に対して，以下の処理を行なった (ただし，サイト C は，他のサイトに比較して記事数が少ないため，頻度 5 以上 10 未満，および，10 以上，の二種類の分布とした).

最初に，英語単語列の品詞を Charniak parser(注4) で調べ，次の簡単な表現を満たすものだけを抽出した.

- W_1 = [形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞] ∗ 名詞
- W_2 = ([形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞]+,) ∗ [形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞] + *and* [形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞] ∗ 名詞

ここで，∗ は 0 回以上の繰り返し，+ は 1 回以上の繰り返しを表す．次に，単語列の上で包含関係にある単語列同士をグルーピングし，英語タームグループを作成した．そして，ある英語タームのグループについて，その要素となる英語ターム t_E が任意の日本語訳語候補に対して持つ ϕ^2 統計値 $\phi^2(t_E, t_J)$ の最大値を，そのグループの持つ ϕ^2 統計値とみなして，英語タームグループを ϕ^2 統計値の降順に整列した．この整列済み英語タームグループから，

(i) グループの上位から順に評価用英語タームを 100 個選定,

(ii) ϕ^2 統計値の決められた範囲 (1 ～ 0.15 および 0.15 ～ 0.07) から，無作為に評価用英語タームを 100 個ずつ選定,

という合計三種類の評価用英語タームセットを作成した(注5)．ただし，サイト A，および，サイト B については，頻度 5 以上 10 未満，10 以上 20 未満，20 以上の三通りの頻度分布ごとにこれらの評価用英語タームセットを作成し，合計で 9 個のタームセットを作成した．サイト C については，頻度 5 以上 10 未満，10 以上の二通りの頻度分布ごとにこれらの評価用英語タームセットを作成し，合計で 6 個のタームセットを作成した.

また，選定の際には，各新聞記事を参照しながら，冗長部分を持つもの，ある単語列の断片であるものを省き，一般的で訳語が一意に定まらないようなもの，および人名と地名を除いた．その上で，日本語訳語候補に正解訳語が含まれている，いないに関わらず，英語タームが妥当であると判断したものを選定した．ただし，正解である日本語訳語が接頭詞，名詞，動詞の任意列でないものは，今回の実験条件では獲得不能であるため除外した.

評価用英語ターム数の分布を表 4 に示す(注6)．それぞ

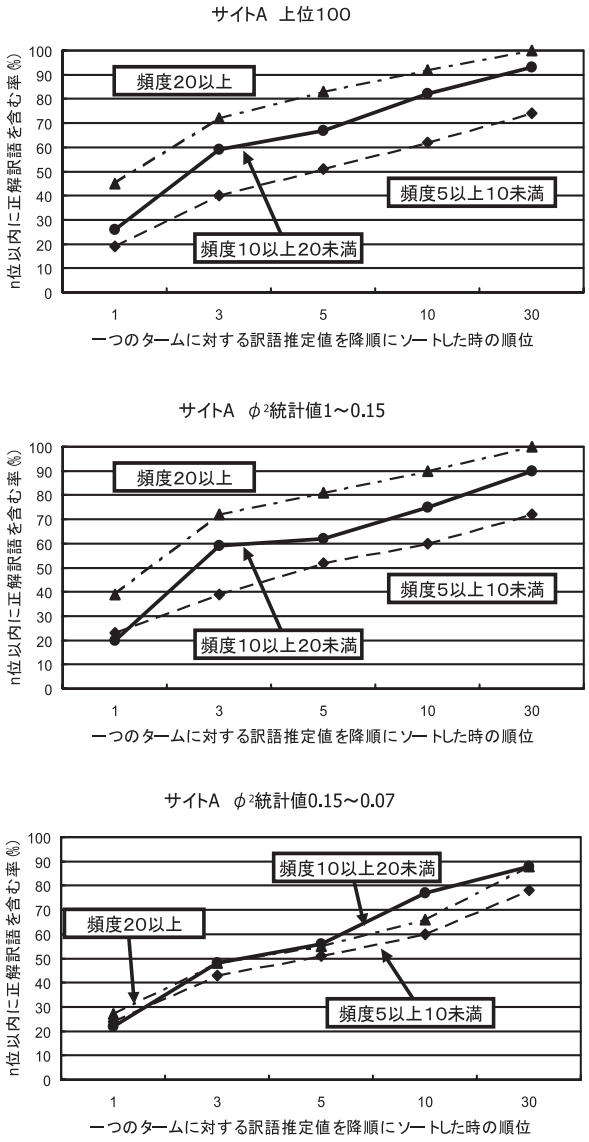


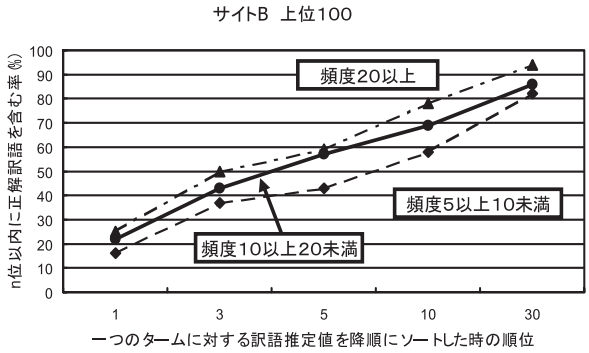
図 2　英語タームの頻度分布 及び ϕ^2統計値の分布 ごとの訳語対応推定性能 (サイト A)

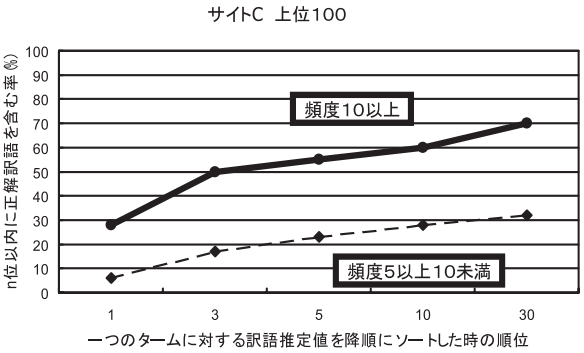
(注3)：http://cl.aist-nara.ac.jp/~taku-ku/software/ prefixspan/

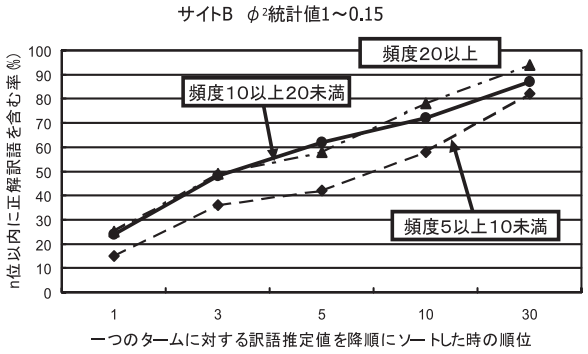
(注4)：http://www.cs.brown.edu/people/ec/

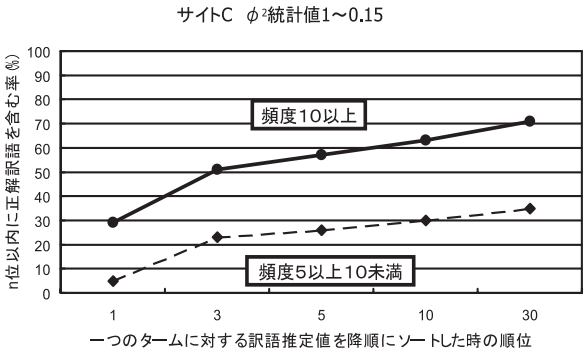
(注5)：いずれも，本稿で用いた翻訳ソフトおよび対訳辞書では訳せないタームから構成される．100 個に満たない場合は可能な限り選定する.

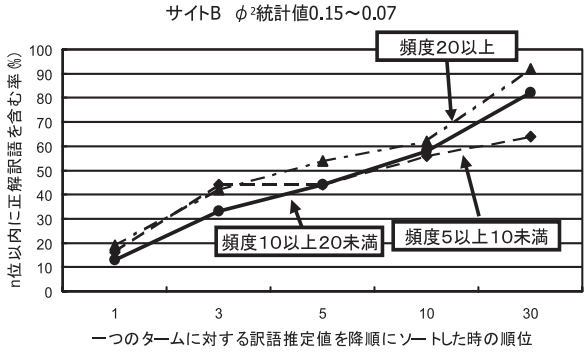
(注6)：実際は, 英語タームのグループ数だが，グループは長いタームの











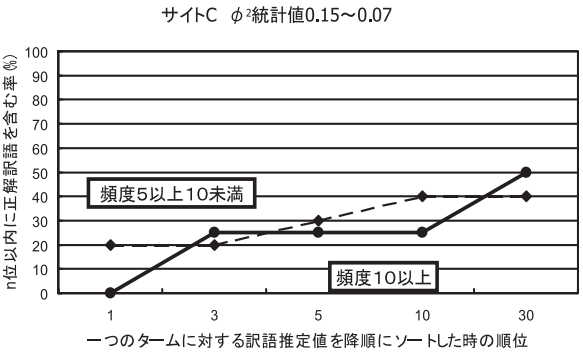


図 3　英語タームの頻度分布 及び ϕ^2統計値の分布 ごとの訳語対応推定性能 (サイト B)

図 4　英語タームの頻度分布 及び ϕ^2統計値の分布 ごとの訳語対応推定性能 (サイト C)

れのサイトに対して，翻訳ソフト(注7)で翻訳に成功した英語ターム数 (「MT」の欄)，対訳辞書(注8)のエントリに含まれる英語ターム数 (「辞書」の欄)，人手で選定した英語ターム数 (「人手」の欄 — 評価用実験で使用した英語ターム)，および，上記の理由により除外した英語ターム数 (「除外」の欄) を示す．ただし，翻訳ソフトで翻訳できる英語タームおよび対訳辞書のエントリに含まれる英語タームの間には重複があり得る．また，「ϕ^2統計値 1 ～ 0.15」および「ϕ^2統計値 0.15 ～ 0.07」の列には，英語タームの ϕ^2統計値 について，それぞれの範囲内で無作為に評価用英語タームを選定した場合のターム数の分布を示す．また，「ϕ^2 統計値 上位 100」の列には，ϕ^2統計値 の降順に，評価用英語タームを 100 個選定した場合のターム数の分布を示す．ここで，「ϕ^2 統計値 上位 100」における ϕ^2 統計値の値は，サイト A において，$\phi^2 > 0.36$，サイト B において，$\phi^2 > 0.16$，サイト C において，$\phi^2 > 0$ となっていた．実際に選定した評価用ターム組の例を表 5 に示す．

表 4 において，例えば，サイト A に対して ϕ^2 統計値が 1～0.15，頻度分布が 5 以上 10 未満の英語タームに注目すると，総数は 1,357 個，対訳辞書のエントリに含まれたものが 285 個，翻訳ソフトで訳せたものが 58 個，対訳辞書のエントリに含まれず翻訳ソフトでも訳せず，訳語対応の獲得対象として判定したターム数は 148 個，対訳辞書のエントリに含まれず翻訳ソフトでも訳せないが，訳語対応の獲得対象とは判定されなかったターム数が 868 個となる．表 4 から分かるように，英語タームの選定においては，「除外」と判定されるタームの割合が大きい．現在の技

部分単語列を含んでいるだけなので，ターム数とグループ数はほぼ等しい．
(注7)：オムロン社製「翻訳魂」
(注8)：英辞郎 Ver.37，85 万語

術レベルでは，訳語対応の獲得対象とすべき英語タームを全自動で報道記事が選定することは困難であり，英語ターム選定の行程において人手の介在が不可欠となっている．

4.3 訳語対応推定の性能

前節で選定した各サイトの評価用英語タームについて，訳語対応推定値の上位 n 位以内に正解訳語 (今回の実験では，各英語タームにつき一つだけ) が含まれる英語タームの割合をプロットした結果を図 2〜 図 4 に示す．ただし,「ϕ^2 統計値上位 100」,「ϕ^2統計値 1 ～ 0.15」,「ϕ^2統計値 0.15 ～ 0.07」の各々の英語タームセットごとにプロットをまとめた．また，訳語対応推定結果の例を表 6 に示す．表内の太字部分が正解日本語訳語である．

全体としては，英語タームの頻度が大きい方が，訳語対応推定の性能が高い．ただし,「ϕ^2統計値 0.15 ～ 0.07」では，英語タームの頻度分布の違いの影響はかなり小さくなっている．つまり，訳語対応推定値 (ϕ^2統計値) が十分大きくなければ，訳語対応推定の性能は，英語タームの頻度によらず，ほぼ同等となると言える．

次に，訳語対応推定性能をサイト間で比較すると，特に，サイト C は，サイト A およびサイト B と比較して，低頻度ターム (頻度 5 以上 10 未満) に対する訳語対応推定性能が低くなっている．サイト C の場合，サイト A およびサイト B と比較して，報道記事の数が約 10 分の 1 と少ないために，英語記事に対応する関連日本語記事を十分収集することができず，結果的に正解訳語との共起頻度が小さくなってしまっていると考えられる．また，サイト A とサイト B を比べると，ϕ^2 統計値の上位において，頻度 20 以上のタームに対する訳語対応推定性能の差が顕著である．この原因を分析するために，次に，訳語対応推定値の一位が正解訳語とならない場合の誤りの内訳を調査した．誤りの原因は主に次の三種類に分類される．

(1) 正解訳語の部分単語列が同等もしくはそれ以上の訳語対応推定値を持つ．
(2) 報道記事中における関連タームが同等もしくはそれ以上の訳語対応推定値を持つ．
(3) 正解訳語との共起頻度が小さい．

(1) の例としては，表 6 の “intellectual property rights” の「知的財産」と**「知的財産権」**などがある．(2) の例としては，“West Nile virus” の「蚊」と**「西ナイルウイルス」**などがある．また (3) は，関連記事対検索が失敗した場合や，もともと日本語関連記事において正解訳語が出現しない場合に起こる．

サイト A およびサイト B において,「ϕ^2 統計値上位 100」のタームセットにおける頻度分布ごとに誤り原因の内訳を求めた結果を表 7 に示す．両サイトの最も顕著な違いとしては，頻度 20 以上のタームセットにおいて,「正解訳語のとの共起頻度が小さい」が占める割合の違いが挙げられる．サイト A ではこの割合が 0 となるのに対して，サイト B ではこの割合が 15%と大きい．これは，サイト A とサイト B では，特に，日本語記事の文体等の特性が異なっており，サイト B では日本語関連記事において正解訳語が出現しないということが一定の割合で起こるためであると考えられる．

表 6　訳語対応推定例

順位	日本語訳語候補	訳語対応推定値
英語ターム	Environment Ministry	
1	**環境省**	**0.541**
2	国定公園	0.099
3	鳥獣保護	0.079
英語ターム	West Nile virus	
1	蚊	0.833
2	ナイルウイルス	0.714
2	**西ナイルウイルス**	**0.714**
英語ターム	intellectual property rights	
1	知的財産	0.094
2	**知的財産権**	**0.079**
3	財産権	0.073

表 7　訳語候補順位付けの誤り原因の分析

サイト	頻度	誤り数	誤り原因の内訳 (%)		
			(1)	(2)	(3)
A	5〜10	81	30	27	43
	10〜20	78	37	47	15
	20 以上	57	44	56	0
B	5〜10	84	33	44	23
	10〜20	78	23	59	18
	20 以上	75	24	61	15

表 8　小規模新聞記事データセットにおけるタームの文書頻度および記事の日数

データセット	タームの文書頻度		日数	
	$df(t_E)$	$df(t_E, t_J)$	英記事	日記事
頻度=10, 13.6 日	14.9	9.1	13.6	21.9
頻度=10, 20 日	14.9	9.1	21.0	78.7
頻度=10, 200 日	14.9	9.1	200	581
頻度=70, 600 日	37.4	24.9	600	872
full	53.9	35.6	935	941

4.4 記事数と訳語対応推定性能の相関

次に，本節では，報道記事における記事数と訳語対応推定性能の相関を評価した結果について述べる．まず，サイト A における出現頻度 10 以上の評価用英語タームのうち，ϕ^2統計値 で降順に上位 100 個を選択し，新たに評価用英語タームセットとした．まず，このタームセットについて，訳語対応推定値の上位 n 位以内に正解訳語が含まれるタームの割合をプロットした結果を，図 5 中の「全記事」に示す．また，図 5 中のその他のプロットでは，報道記事の日数を減少させて，疑似的に記事数を減少させ，ターム頻度を減らせた場合の訳語対応推定性能を示す．ただし,「頻度=x, y 日」というラベルのプロットにおいては,

（1） 評価用英語タームおよび正解日本語訳語の共起頻度を x に固定．
（2） 報道記事の日数は y 以上．

という条件を満たすように報道記事データセットを作成する．また，これらのデータセットにおけるタームの文書頻

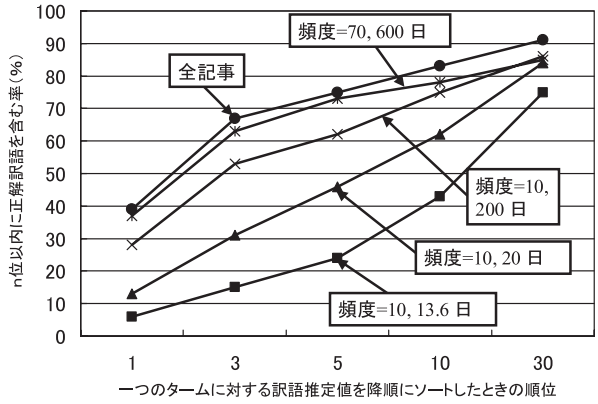


図 5　記事数と訳語対応推定性能の相関

度および記事の日数の内訳を表 8 に示す．図 5 から分かるように，報道記事の規模が小さくなるほど，訳語対応推定の性能は低下する．特に，最も小規模な記事データセットにおいては,「全記事」と比較して，訳語対応推定の性能が大幅に低下することが分かる．

5. おわりに

本稿では，日英関連報道記事からの訳語対応推定のタスクにおいて，英語タームの出現頻度と, 訳語対応推定性能の相関を評価し, 英語タームの頻度が大きいほど，高い訳語対応推定性能が達成できることが分かった．今後は，ウェブ検索エンジンにより収集した日英非対訳文書より得られる訳語候補の順位付け [4] と，日英報道記事から得られる順位付けの統合を行い，訳語対応推定の性能の改善を試みる予定である．

文　献

[1] P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. 17th COLING and 36th ACL*, pp. 414–420, 1998.

[2] 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, pp. 425–428, 2003.

[3] 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.

[4] 木田充洋, 宇津呂武仁, 日野浩平, 佐藤理史. 日英二言語文書を用いた訳語対応推定: ウェブ上の非対訳文書を用いた訳語候補順位付け. 情報処理学会研究報告, Vol. 2004, No. (2004–NL–162), 2004.

[5] Y. Matsumoto and T. Utsuro. Lexical knowledge acquisition. In R. Dale, H. Moisl, and H. Somers, editors, Handbook of Natural Language Processing, chapter 24, pp. 563–610. Marcel Dekker Inc., 2000.

[6] J. Pei, J. Han, B. Mortazavi-Asl, and H. Pinto. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. Inter. Conf. Data Mining*, pp. 215–224, 2001.

[7] R. Rapp. Identifying word translations in non-parallel texts. In *Proc. 33rd ACL*, pp. 320–322, 1995.

[8] T. Utsuro, K. Hino, M. Kida, S. Nakagawa, and S. Sato. Integrating cross-lingually relevant news articles and monolingual Web documents in bilingual lexicon acquisition. In *Proc. 20th COLING*, 2004. (to appear).

[9] T. Utsuro, T. Horiuchi, T. Hamamoto, K. Hino, and T. Nakayama. Effect of cross-language IR in bilingual lexicon acquisition from comparable corpora. In *Proc. 10th EACL*, pp. 355–362, 2003.