日英関連報道記事を用いた訳語対応推定

宇津呂 武仁[†]　日野　浩平[††]
堀内　貴司[†††]　中川　聖一[††††]

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．本論文では，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獲得する手法を提案する．本論文の手法には，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た翻訳知識獲得の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．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，関連記事組を用いて二言語間の訳語対応を推定する．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を適用し，評価実験において文脈ベクトルを用いる方法と比較し，この方法が有効であることを示す．

キーワード: 機械翻訳，訳語対応推定，対訳コーパス，コンパラブルコーパス，言語横断情報検索，対訳辞書

Estimating Bilingual Term Correspondences from Relevant Japanese-English News Articles

Takehito Utsuro[†] , Kohei Hino[††] , Takashi Horiuchi[†††]
and Seiichi Nakagawa[††††]

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. We experimentally show that the proposed method is effective in estimating bilingual term correspondences from cross-lingually relevant news articles.

KeyWords: *machine translation, estimating bilingual term correspondences, parallel corpus, comparable corpus, cross-language IR, bilingual lexicon*

1　はじめに

近年，コーパスを利用した機械翻訳の研究においては，翻訳システムに不足している翻訳知識を人手で増強していく際のコストを軽減する目的で，対訳コーパスやコンパラブルコーパス等の多言語コーパスから様々な翻訳知識を獲得する手法の研究が行なわれてきた (Matsumoto and Utsuro 2000)．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．通常，対訳コーパスからの獲得 (例えば，(Gale and Church 1991)) 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，(Rapp 1995; Fung and Yee 1998)) 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，対訳コーパスと比較すると，片方の言語のコーパス中のタームや表現の訳がもう一方の言語のコーパスに出現する可能性が相対的に低いため，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．

そこで，本論文では，翻訳知識獲得の目的において，人手で整備された対訳コーパスよりも利用可能性が高く，一般のコンパラブルコーパスよりも翻訳知識の獲得が容易である情報源として，日英二言語で書かれた報道記事に着目する．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．そこで，本論文では，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，分野固有の人名・地名・組織名などの固有名詞 (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．本論文のアプローチは，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．

本論文の翻訳知識獲得のアプローチにおいて，日英関連報道記事から翻訳知識を獲得する

† 京都大学情報学研究科知能情報学専攻, Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University
†† NTT データテクノロジー株式会社, NTT Data Technology Corporation
††† 日立製作所, Hitachi Ltd.
†††† 豊橋技術科学大学工学部情報工学系, Department of Information and Computer Sciences, Toyohashi University of Technology

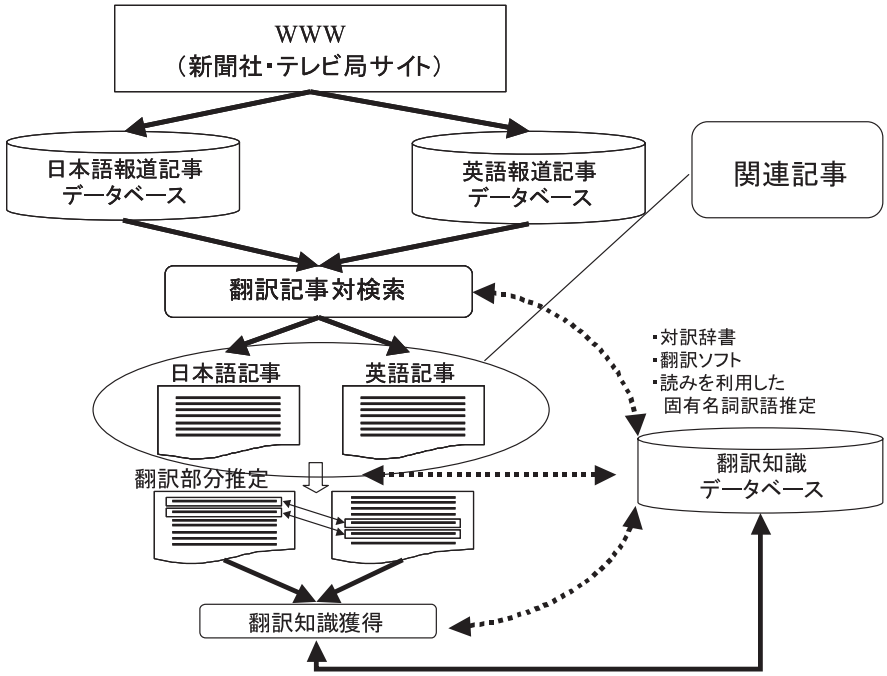


図 1　日英関連報道記事からの翻訳知識獲得のプロセス

プロセスの一般的な流れを図 1 に示す．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．この際には，既存の対訳辞書，翻訳ソフトの翻訳知識を利用することにより，日本語記事と英語記事の間の関連性を測定する．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，従来のコンパラブルコーパスからの訳語対獲得のアプローチにおいては，原理的には，コンパラブルコーパスに出現する全ての日本語タームおよび英語タームの組を訳語対応の候補としていた．一方，本論文のアプローチでは，予備調査の結果 (Utsuro，Horiuchi，Hamamoto，Hino and Nakayama 2003; 堀内，日野，浜本，中山，宇津呂 2003) をふまえて，関連報道記事の組において共起した日本語ターム，および，英語タームの組を収集し，これを訳語対応の候補としており，この点が特徴的である[1]．ただし，本論文で述べる手法の範囲では，現在のところ，関連記事中で内容的に対応する翻訳部分の推定は行なっておらず，関連記事対全体から訳語対応を推定している．また，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

1 予備調査の結果 (Utsuro et al. 2003; 堀内他 2003) においては，関連報道記事の組において共起した日本語ターム，および，英語タームの組を訳語対応の候補とすることにより，不要な訳語対応の候補を大幅に削減できることが分かっており，本論文のアプローチが適切であることの裏付けとなっている．

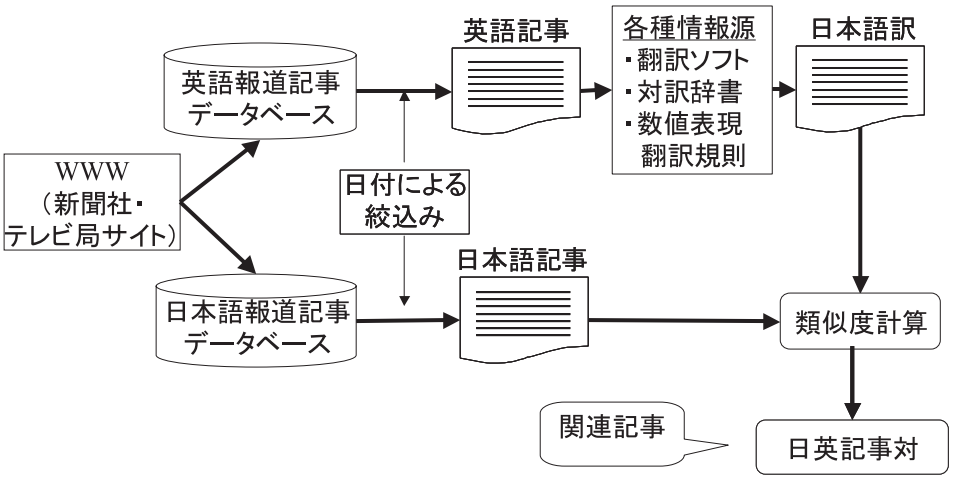


図 2　日英関連報道記事検索のプロセス

を適用し，評価実験を通して，この方法が有効であることを示す．特に，評価実験においては，訳語対応を推定すべき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どのように変化するかを調査し，その相関を評価する．

以下，2節では，翻訳知識獲得のための情報源収集を目的として，言語を横断して，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処理について述べる．次に，3節では，関連記事組の集合から訳語対応を推定する手法について述べる．4節において，実験を通して提案手法の評価を行ない，5節において，関連研究について詳細に述べる．

2　言語横断関連報道記事検索

本論文の翻訳知識獲得のアプローチにおける言語横断関連報道記事検索の流れを図 2 に示す．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，内容的にほぼ同一の日英記事対は，お互いの日付が前後数日程度の範囲にあるという調査結果 (堀内，千葉，浜本，宇津呂 2002b, 2002a) に基づいて，日付の情報を用いて検索対象の記事を絞りこむ (実際に，評価実験において用いた日英記事間の日付の幅の詳細については，4.1 節で述べる．)．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・数値表現翻訳規則などの情報源を利用して英語記事 d_E を日本語訳に変換する．ここで，言語横断関連報道記事検索の性能において，翻訳ソフト (オムロン社製「翻訳魂」)，対訳辞書 (英辞郎 Ver.37，85 万語)，および，数値表現翻訳規則 (規則数約 300) の三種類の情報源の性能を比較した結果においては，翻訳ソフトが最も高い検索性能を達成した (浜本，中山，日野，堀内，宇津呂 2003)．そこで，本論文の評価実験

においても，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[2]．

次に，英語記事 d_E の日本語訳から日本語訳頻度ベクトル $v_{trJ}(d_E)$ を，また，日本語記事 d_J から日本語頻度ベクトル $v(d_J)$ を，それぞれ作成する．ここでは，日本語形態素解析システム「茶筌」[3] を用いてテキストを形態素列に分割し，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，頻度ベクトルにおいては，接頭詞，名詞，動詞によって構成され，形態素長が 5 以内の形態素列を次元とした[4]．最後に，頻度ベクトル間で余弦類似度を計算し，余弦類似度が下限値以上の記事を関連記事検索結果とする．

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．詳細については 4.1 節で述べるが，本論文の評価実験において対象とした新聞社・テレビ局のサイトは日本国内のもので，掲載される報道記事数は，日本語記事数が英語記事数の 4~6 倍となっている．したがって，検索質問として日本語記事を用いる場合よりも，検索質問として英語記事を用いた場合の方が，関連記事組が収集できる割合が大きい．実際に，検索質問として英語記事を用いた場合には，検索質問の約半数に対して関連記事組が収集できることが分かっている (堀内他 2002b, 2002a)．これらの調査結果をふまえて，本論文では，英語記事を検索質問として関連日本語記事を収集することにより，日英関連記事組を作成する．ここで，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．

$$D_J \quad = \quad \Big\{ d_J \mid \cos(v_{trJ}(d_E), v(d_J)) \geq L_d \Big\}$$

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 (ここで，D_J 中の記事を結合して一つの日本語記事 D'_J を構成するのは，3.1 節において，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なうためである．).

$$RC_{EJ} \quad = \quad \Big\{ \langle d_E, D'_J \rangle \mid D_J \neq \emptyset \Big\} \tag{1}$$

2 (日野，宇津呂，中川 2004b) においては，訳語対応推定の性能において，翻訳ソフトを用いて日英関連報道記事を検索した結果，および，対訳辞書を用いて日英関連報道記事を検索した結果を比較しているが，ここでも，翻訳ソフトを用いた方が高い性能となっている．

3 http://chasen.naist.jp/hiki/ChaSen/

4 各記事の頻度ベクトルの次元としては，3.2 節で述べる文単位の文脈頻度ベクトルの次元と同じものを用いている．予備調査の結果，文脈頻度ベクトルを用いた訳語対応推定においては，一形態素のみを次元とした文脈頻度ベクトルでは不十分であるが，5 形態素長以内の形態素列を次元としておけば，周囲の文脈として必要な表現がほぼ含まれることが分かっている．これは，5 形態素を越える長さの形態素列を用いないと，周囲の文脈の特性を表現しきれない，ということが極めて稀であるからである．一方，関連記事検索の性能を評価した予備調査の結果においては，名詞および動詞の一形態素のみを次元とした場合と，5 形態素長以内の形態素列を次元とした場合との間の性能差はわずかであった．以上をふまえて，本論文では，頻度ベクトルの次元としては，形態素長が 5 以内の形態素列を用いる．

表 1　2 × 2 分割表

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

3　日英関連報道記事における訳語対応の推定

本論文では，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．本節では，関連記事組の集合を疑似的な対訳コーパスとみなす場合の方法を 3.1 節で，関連記事組の集合をコンパラブルコーパスとみなす場合の方法を 3.2 節で，それぞれ説明する．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．

3.1　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，対訳コーパスからの訳語対応推定の場合と同様に，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 (Matsumoto and Utsuro 2000) が適用可能である．これら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，dice 係数，対数尤度比がほぼ同程度の性能となり，相互情報量はやや劣るという結果が得られた．そこで，本論文では，t_E と t_J の統計的相関を測定する尺度としては，ϕ^2 統計を用いることとし，これを訳語対応推定値 $corr_{EJ}(t_E, t_J)$ とする．具体的には，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 $df(t_E, t_J)(= a$ とする)，t_E のみが含まれ t_J が含まれない記事組数 $df(t_E, \neg t_J)(= b$ とする)，t_J のみが含まれ t_E が含まれない記事組数 $df(\neg t_E, t_J)(= c$ とする)，t_E も t_J も含まれない記事組数 $df(\neg t_E, \neg t_J)(= d$ とする) を用いて表 1 の 2 × 2 分割表を構成する．この 2 × 2 分割表を用いると，t_E と t_J の ϕ^2 統計は以下で与えられる．

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

3.2　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 についての文単位の文脈頻度ベクトルを求め，これらの文脈頻度ベクトル間の類似性を用いて t_E と t_J の訳語対応を推定する．具体的には，前節で述べたように，英語記

事 d_E に対する日本語訳頻度ベクトルを $v_{trJ}(d_E)$ として，d_E において t_E が出現する文の日本語訳の頻度ベクトルを $v_{trJ}(d_E)$ から求め，これを加算して，t_E に対する文単位の文脈頻度ベクトル $cv_{trJ}(t_E)$ を構成する．同様に，日本語記事 d_J を集めた記事集合において t_J が出現する文について，それらの頻度ベクトルを加算することにより，t_J に対する文単位の文脈頻度ベクトル $cv(t_J)$ を構成する．そして，この文脈頻度ベクトル間の余弦 $\cos(cv_{trJ}(t_E), cv(t_J))$ を $corr_{EJ}(t_E, t_J)$ とする．

4　実験および評価

4.1　言語横断関連報道記事検索

国内の新聞社等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．また，表 2 には，言語横断関連報道記事検索の性能の評価のために用いる評価用日英記事対数も示す．ここで，本論文では，報道内容がほぼ同一の日英記事対のことを「同一内容」の記事対とよび，報道内容は同一ではないが，記事として密接に関連している日英記事対 (例えば，事件発生に関する報道記事に対して，犯人逮捕に関する続報記事など) のことを「関連話題」の記事対とよぶ．

次に，2 節で述べたように，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．評価用日英記事対のうちの英語記事を検索質問として，日本語記事を検索した場合の適合率・再現率の変化をプロットしたものを図 3 に示す．ここで，評価用 (同一内容または関連話題) 記事対の集合を DP_{ref}，記事間類似度の下限値を L_d とすると，この場合の適合率・再現率の定義は，

$$\text{適合率} = \frac{|\{\langle d_E, d_J\rangle \mid \langle d_E, d_J\rangle \in DP_{ref}, \cos(d_E, d_J) \geq L_d\}|}{|\{d_E \mid \exists d'_J, \langle d_E, d'_J\rangle \in DP_{ref}, \exists d_J \cos(d_E, d_J) \geq L_d\}|}$$

$$\text{再現率} = \frac{|\{\langle d_E, d_J\rangle \mid \langle d_E, d_J\rangle \in DP_{ref}, \cos(d_E, d_J) \geq L_d\}|}{|\{\langle d_E, d_J\rangle \mid \langle d_E, d_J\rangle \in DP_{ref}\}|}$$

となる．

また，表 3 には，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示す．表 3 においては，各サイトについて用いた日付の幅もともに示す．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．また，図 3 においても，類似度下限 L_d が 0.4 や 0.5 の場合は，再現率が大きく低下している．ここで，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．

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)	評価用記事対数	
						同一内容	関連話題
サイト A	英語	935	23064	24.7	3228.9	28	31
	日本語	941	96688	102.8	837.7		
サイト B	英語	935	14587	15.6	3302.6	28	82
	日本語	941	81652	86.8	867.9		
サイト C	英語	935	1553	1.6	1368.6	24	33
	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	0.3	0.4	0.5
サイト A	日付幅 (日)	± 2		
	英語記事数	6073	2392	701
	日本語記事数	12367	3444	882
	日本語記事数 (重複あり)	16507	3840	918
サイト B	日付幅 (日)	± 2		
	英語記事数	4316	1658	396
	日本語記事数	8108	2349	499
	日本語記事数 (重複あり)	11451	2694	523
サイト C	日付幅 (日)	± 4		
	英語記事数	765	413	159
	日本語記事数	1918	673	192
	日本語記事数 (重複あり)	2406	766	203

4.2　英語・日本語訳語組候補の条件

本論文では，実装の都合上，英語・日本語間で訳語組候補となるタームに対して，タームを構成する単語もしくは形態素の数に上限を設け，さらに，タームを構成する単語もしくは形態素の品詞にも制限を設ける．まず，タームを構成する単語もしくは形態素の数については，上限を 5 とする．この条件により，次節で選定される評価用英語ターム (およびその正解日本語訳語) については，構成単語数あるいは構成形態素数が 5 を越えるものは除外される．また，英語タームとしては名詞句を対象とすることとし，Charniak parser[5] を用いて各英語単語の品詞付けを行い，英語単語列として以下の表現を満たすものだけを対象とする (ただし，$*$ は 0 回以上の繰り返し，$+$ は 1 回以上の繰り返しを表す．).

- $W_1 =$ [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

5 http://www.cs.brown.edu/people/ec/

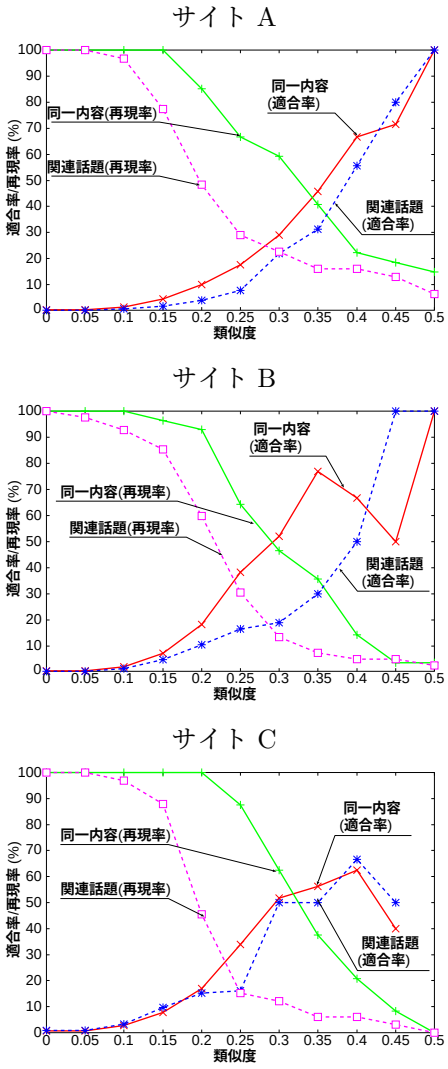


図 3　日英関連記事検索の適合率・再現率 (記事間類似度 $\geq$ L_d)

日本語タームについても，名詞句相当のものを対象とする．具体的には，接頭詞，名詞，動詞によって構成される形態素列を対象とする (構成要素に動詞を含めたのは，連用形の名詞用法に対応するためである．ただし，現時点では，活用形の照合は行なっていないため，日本語タームの候補として適切でないものも混入している．).

さらに，2 節の (1) 式において，英日関連記事組を集めて構成した集合 RC_{EJ} 中の関連記事組 $\langle d_E, D'_J \rangle$ において，英語ターム t_E と日本語ターム t_J が共起する記事組数 $df(t_E, t_J)$ に下限

を設け，これを 2 以上とする．

4.3　評価用英語タームの選定

本論文の訳語対応推定の評価実験の範囲では，訳語対応推定の対象とする英語タームを自動抽出することは行わず，訳語対応推定の評価用英語タームを人手で選定しておき[6]，これらに対して日本語訳語候補を自動抽出し，日本語訳語候補の順位付け性能の評価を行った．特に，本論文では，既存の翻訳ソフト (オムロン社製「翻訳魂」) によって翻訳することができず，対訳辞書 (英辞郎 Ver.37，85 万語) にも存在しない英語タームを評価用英語タームとして選定した．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 (Pei，Han，Mortazavi-Asl and Pinto 2001)[7]を用いて頻度 5 以上の単語列の頻度を測定した．そして，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 (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．) で単語列集合を分割し，それぞれの集合に対して，以下の手順によって評価用英語タームを選定した．

(1)　英語タームグループの作成
(2)　ϕ^2 統計値を用いた英語タームグループの整列
(3)　評価用英語タームの選定

以下，これらの手順の詳細について順に述べる．

4.3.1　英語タームグループの作成

前節で述べた通り，本論文では，英語タームとしては名詞句を対象とする．そこで，まず，前節の制限を満たす英語単語列を抽出する．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．このとき，多くの場合，一つの英語タームグループ中において，適切な英語タームとして認定すべき単語列は高々一つ程度であるので，実質的なターム数はタームグループ数とほぼ等しい．このことをふまえて，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ごとの英語タームの内訳を表 4(a) に示す (ただし，サイト C については，頻度 5 以上 10 未満，および，10 以上，の二種類の出現頻度分布とする)．英語タームの内訳は，翻訳ソフトで翻訳に成功した英語ターム数 (「MT」の欄)，対訳辞書に存在する英語ターム数 (「辞書」の欄)，および，翻訳ソフトによって翻訳することができず，対訳辞書にも存在しない英語ターム数 (「その他」の欄) によって示す．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に存在する英語タームの間には重複があり得る．

6 既存のターム抽出技術を用いることにより，一定レベルの性能で英語タームを抽出することは可能である．本論文の訳語対応推定の枠組において，英語ターム自動抽出の技術を併用すれば，訳語組を全自動で獲得する一連の流れの性能を評価することができると考えられる．

7 http://chasen.org/~taku/software/prefixspan/

表 4　評価用英語ターム数の分布

(a) 全体

サイト	頻度	5∼10	10∼20	20 以上
A	MT	117	158	531
	辞書	1391	1718	2507
	その他	4423	3483	2786
	総数	5931	5359	5824
B	MT	104	214	791
	辞書	1098	1367	1968
	その他	3105	2364	1868
	総数	4292	3835	4167
C	MT	103	164	
	辞書	226	205	
	その他	313	152	
	総数	585	424	

(b) ϕ^2統計値 ごとの分布

		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		
サイト	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	58	82	289	32	37	147	38	110	103
	辞書	285	407	727	229	304	570	73	166	157
	評価用	148	116	131	51	48	56	100	100	100
	除外	866	671	687	800	684	618	199	75	66
	総数	1357	1276	1834	1112	1073	1391	397	381	360
B	MT	87	124	377	28	57	226	87	128	95
	辞書	216	321	590	203	236	452	216	333	218
	評価用	104	71	102	25	45	26	100	100	100
	除外	669	476	462	570	432	418	673	487	306
	総数	1048	922	1298	808	740	995	1048	977	668
C	MT	75	114		22	43		103	164	
	辞書	147	125		46	60		226	205	
	評価用	43	35		10	4		57	40	
	除外	158	68		54	33		256	112	
	総数	379	275		123	115		585	424	

4.3.2　ϕ^2 統計値を用いた英語タームグループの整列

次に，ある英語タームグループについて，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．なお，詳細は 4.4 節で述べるが，予備実験 (日野他 2004b) において，3.1 節の ϕ^2 統計を用いる方法と，3.2 節の文脈ベクトルを用いる方法を比較した結果では，ϕ^2 統計を用いた方法の方が高い性能であった．そこで，本論文では，ϕ^2 統計値の降順に整列した英語タームグループを用いて評価用英語タームを選定することとした．

4.3.3　評価用英語タームの選定

この整列済み英語タームグループのうち,「その他」に分類される英語タームグループを人手で選別し，以下の個数の評価用英語タームを選定して，合計三種類の評価用英語タームセットを作成した．

(i)　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 個ずつ選定した．ただし，100 個に満たない場合は可能な限り選定する．

(ii)　上位の英語タームグループに含まれる英語タームから順に評価用英語タームを 100 個選定した．

ただし，選別の際には，各新聞記事を参照しながら，冗長部分を持つもの，別の単語列の断片であるもの，一般的で訳語が一意に定まらないようなもの，および，人名と地名を除外した．その上で，日本語関連記事から収集した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[8] [9]．この手順から分かるように，「その他」に分類される英語タームは，人手による選定の際に，訳語対応推定対象としては適切でないと判断して除外したもの，訳語対応推定対象として適切であり，評価用英語タームとして選定されたもの，および，人手による選別を受けないまま残されたものの三種類のタームから構成される．

この結果，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たため，合計で 9 個のタームセットとなった．また，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たため，合計で 6 個のタームセットとなった．これらのタームセットにおける英語ターム数の内訳を表 4(b) に示す[10]．英語ターム数の内訳は，各タームセットについて，翻訳ソフトで翻訳に成功した英語ターム数 (「MT」の欄)，対訳辞書に存在する英語ターム数 (「辞書」の欄)，人手で選定した英語ターム数 (「評価用」の欄 — ただし，100 個を超える場合には，実際の評価実験において使用したのは 100 個のみ)，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) によって示す [11]．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に存在する英語タームの間には重複があり得る．実際に選定した評価用ターム組の例を表 5 に示す．

表 4(b) 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 866 個となってい

8 この条件により，本論文の訳語対応推定の評価は，各サイトから収集した日英関連記事組において，正解の日本語訳語がどの程度の割合で含まれているかを考慮した評価となっていると言える．実際に，正解の日本語訳語が含まれる度合はサイトによって異なっており，その詳細については，次節で考察する．

9 厳密には，正解である日本語訳語が前節の日本語タームの条件 (接頭詞，名詞，動詞によって構成され，形態素長が 5 以内の形態素列) を満たさない場合には，訳語候補の日本語タームとすることができない．このような場合には，正解日本語訳語との間の訳語対応推定が不可能であるため，評価用英語タームとしては選定しなかった．

10 実際は，英語タームグループ数だが，上述の通り，英語ターム数と英語タームグループ数はほぼ等しい．

11 具体的には，「ϕ^2統計値 1 ～ 0.15」および「ϕ^2統計値 0.15 ～ 0.07」の「評価用」の欄には，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を示し，「ϕ^2 統計値 上位 100」の「評価用」の欄には，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を示す．

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

る．表 4から分かるように，本論文における評価用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くなっている．

4.4　訳語対応推定の性能

本節では，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 (本論文の実験では，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することにより，訳語対応推定の性能を評価する．

まず，サイト A について，頻度 10 以上の評価用英語タームを用いて以下の二種類のタームセットを作成し，3.1 節の ϕ^2 統計を用いる方法と，3.2 節の文脈ベクトルを用いる方法を比較した．

i)　評価用英語タームのうち ϕ^2 統計値の上位 100 タームを集めたタームセット,

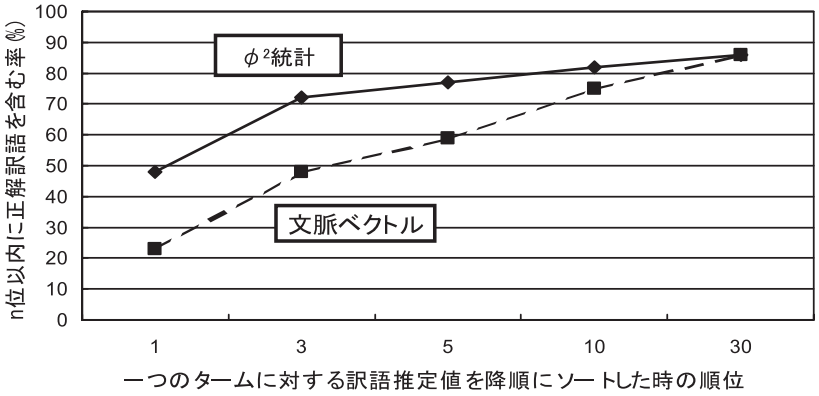
ii)　ϕ^2 統計値の上位 1000 タームグループから無作為に評価用英語タームを 100 ターム選定して作成したタームセット.

セット i) の選定方法では ϕ^2 統計を用いる方法に有利になる可能性があるため，別途，セット ii) を用いた評価も行なった．この結果を図 4 に示す．この結果から分かるように，いずれのセットにおいても，ϕ^2 統計を用いる方法の方が高い性能を示すことが分かる．

次に，サイト A，B，および，C の各サイトについて，表 4(b) の各タームセットに対する訳語対応推定性能を評価した結果を図 5∼ 図 7 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，表 5 の評価用英語タームに対する訳語対応推定結果の抜粋を表 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

(1) 評価用英語タームのうち ϕ^2 統計値の上位 100 ターム



(2) ϕ^2 統計値の上位 1000 タームグループから無作為に評価用英語タームを 100 ターム選定

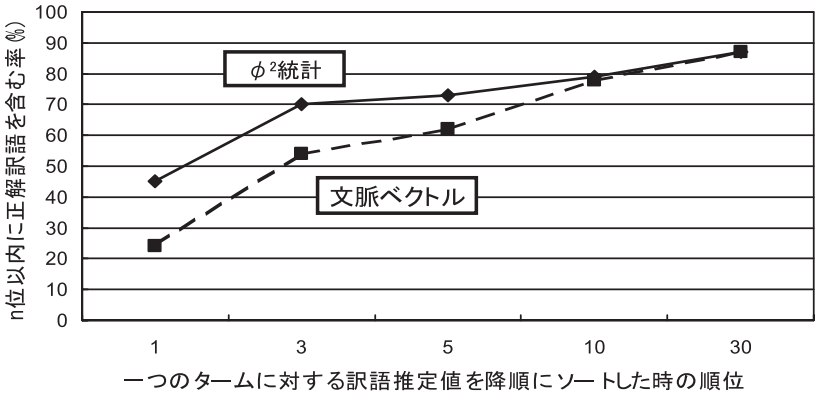


図 4　訳語対応推定手法の比較 (サイト A，頻度 10 以上の評価用英語ターム)

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

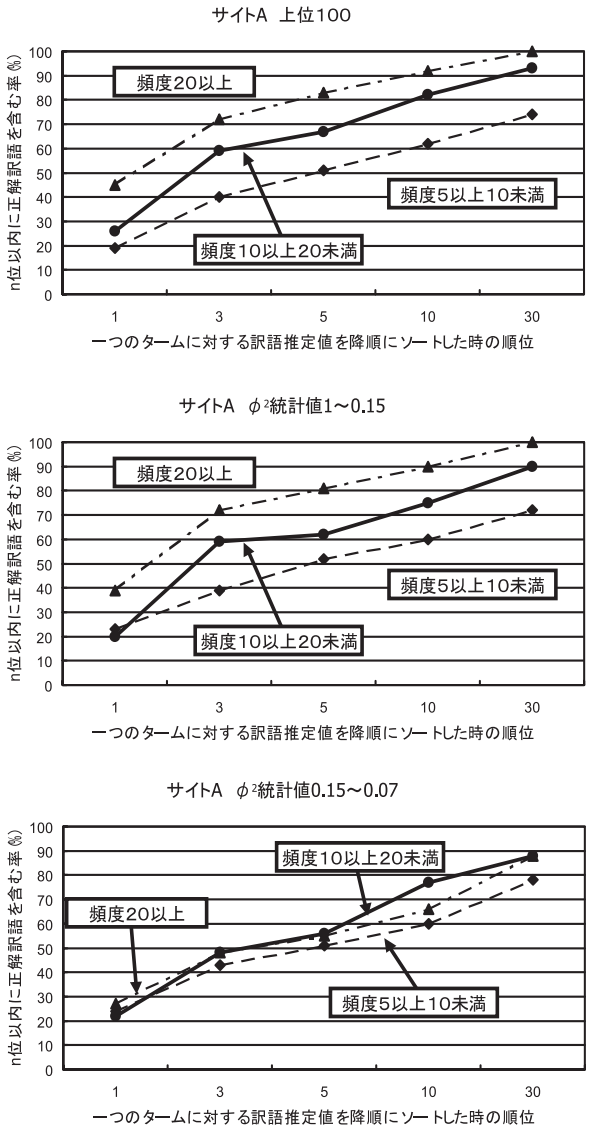


図 5　英語タームの頻度分布 及び ϕ^2統計値の分布 ごとの訳語対応推定性能 (サイト A)

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)　単語列として，正解訳語との間で包含関係にある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．

(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．

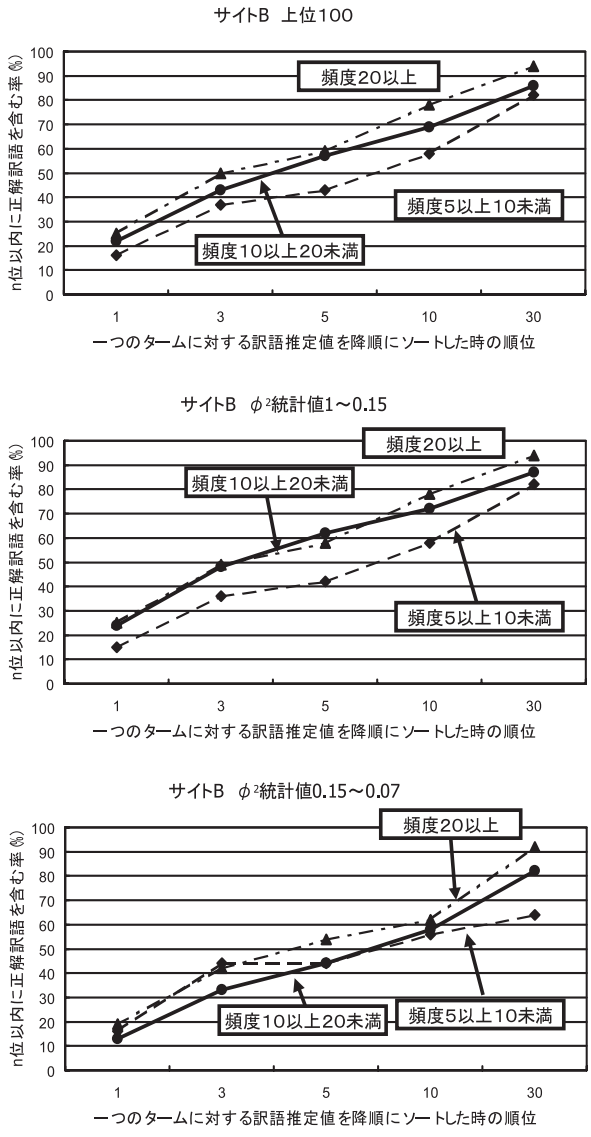


図 6　英語タームの頻度分布 及び ϕ^2統計値の分布 ごとの訳語対応推定性能 (サイト B)

(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 “High Public Prosecutors Office” の「大阪高検」と**「高検」**，“intellectual property rights” の「知的財産」と**「知的財産権」**，“special structural reform zones” の「構造改革特区推進」と**「構造改革特区」**などがある．(2) の例としては，“Japanese Consulate General” の「連行事件」と**「日本総領事館」**，“diesel-powered vehicles” の「浄化装置」と

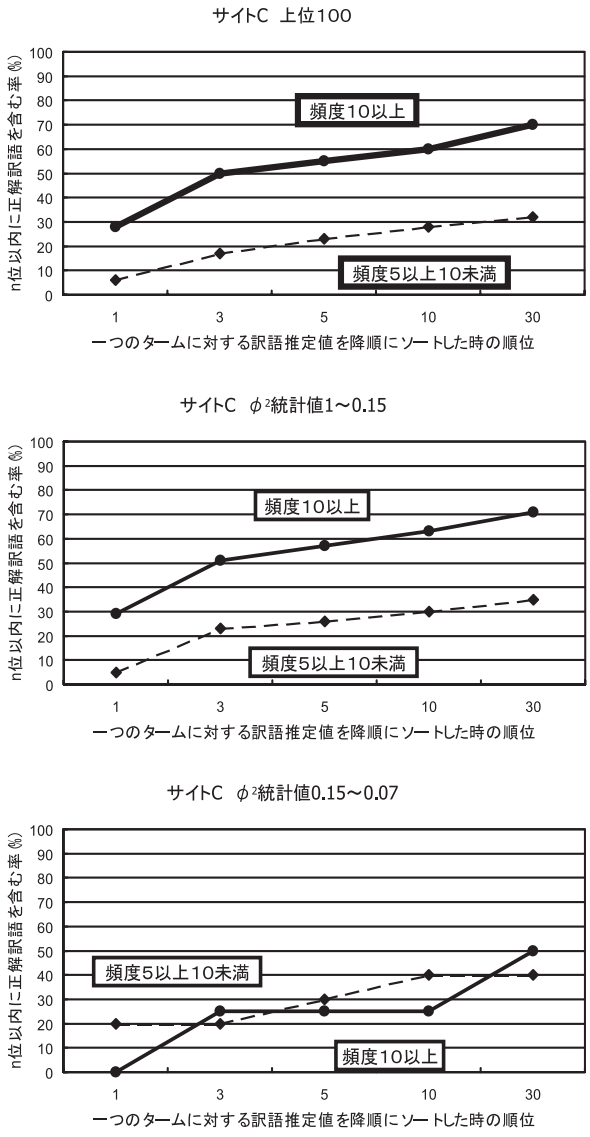


図 7　英語タームの頻度分布 及び ϕ^2統計値の分布 ごとの訳語対応推定性能 (サイト C)

「ディーゼル車」，“fertilized eggs ” の「ES 細胞」と**「受精卵」**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．

さらに，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．両サイト間の最も顕著な違いとし

表 6　訳語対応推定例 (**太字: 正解訳語**)

英語ターム	順位	日本語訳語候補	訳語対応推定値
High Public Prosecutors Office	1	大阪高検	0.640
	2	公安部長	0.450
	2	三井環	0.450
	4	登録免許税	0.360
	5	**高検**	**0.290**
Environment Ministry	**1**	**環境省**	**0.542**
	2	国定公園	0.099
	3	鳥獣保護	0.079
Japanese Consulate General	1	連行事件	0.521
	2	中国・瀋陽	0.507
	3	亡命者連行事件	0.497
	4	亡命者	0.482
	5	瀋陽	0.393
	6	**日本総領事館**	**0.389**
diesel-powered vehicles	1	浄化装置	0.520
	2	粒子状物質	0.408
	3	**ディーゼル車**	**0.382**
Japan Coast Guard	**1**	**海上保安庁**	**0.503**
	2	巡視	0.394
	3	海保	0.382
fertilized eggs	1	ES 細胞	0.500
	2	**受精卵**	**0.333**
	2	不妊治療	0.333
Tokyo District Public Prosecutors Office	**1**	**東京地検**	0.443
	2	東京地検特捜部	0.378
	3	地検特捜部	0.343
Aum Supreme Truth	**1**	**オウム真理教**	**0.467**
	2	松本被告	0.432
	3	こと松本智津夫	0.410
intellectual property rights	1	知的財産	0.095
	2	**知的財産権**	**0.080**
	3	財産権	0.073
special structural reform zones	1	構造改革特区推進	0.457
	2	**構造改革特区**	**0.349**
	3	構造改革特区推進本部	0.321

ては，頻度 20 以上のタームセットにおいて,「正解訳語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 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	16
	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

5　関連研究

本節では，コーパスを用いて訳語対応等の翻訳知識を獲得する手法に関連する研究のうち，言語横断関連報道記事検索に関する関連研究，および，訳語対応推定に関する関連研究について述べる．

5.1　言語横断関連報道記事検索

2節で述べた言語横断関連報道記事検索の手法に関連する研究として，内容的に対応した二言語文書を収集する手法に関する研究がいくつか行なわれている．二言語文書の種類としては，同一の期間の報道記事を対象として，内容が対応した二言語の記事を収集するという手法がいくつか提案されている．言語を横断して記事の内容の類似性を測定する手法を分類する観点としては，主として，i) 言語を横断する際に用いる対訳情報の情報源の種類，ii) 記事間の類似度を測定する際に，文レベルの対応まで考慮するか否か，という二点が挙げられる．i) に関しては，翻訳ソフト，既存の対訳辞書，あるいは，内容的に対応する既知の二言語文書から学習した翻訳モデル，等の情報が用いられる．また，ii) に関しては，既存の多くの研究においては，文レベルの対応までは考慮せず，文書全体での類似性を測定している．そのような事例としては，例えば，数値表現や名前等の訳語対応を情報源として用いるもの (Takahashi，Shirai and Bond 1997; Xu and Tan 1999)，翻訳システムおよび会社名の対訳辞書を情報源として用いるもの (Matsumoto and Tanaka 2002)，翻訳システムおよび既存の対訳辞書を情報源として用い，両者の性能比較を行なったもの (Collier，Hirakawa and Kumano 1998) などがある．また，(Masuichi，Flournoy，Kaufmann and Peters 2000) は，特許文書を対象として，小規模な対訳文書を初期データとして，ブートストラップにより言語横断情報検索モデルを学習しながら，内容的に対応する二言語文書を収集する手法を提案している．一方，(Hasan and Matsumoto 2001) は，日中二言語間で内容的に対応する文書を収集するタスクにおいて，翻訳ソフトおよび漢字を利用したいくつかの統計量を情報源として用いている．以上の事例においては，いず

れも，文レベルの対応までは考慮せず，記事全体で類似性を測定している．それに対して，(内山，井佐原 2003) は，読売新聞および The Daily Yomiuri という，完全な対訳に近い二言語文書対の収集がある程度期待できる文書集合を対象として，既存の対訳辞書を情報源として，記事中の文の対応まで考慮した日英記事間の類似度を用いて，内容的に対応する記事を収集する手法を提案している．

これらの関連研究と比較すると，本論文で述べた言語横断関連報道記事検索の手法は，i) の，言語を横断する際に用いる対訳情報の情報源の種類に関しては，2節で述べたように，翻訳ソフト，対訳辞書，数値表現翻訳規則の三種類のうち，単独の情報源としては翻訳ソフトを用いている．また，ii) に関しては，文レベルの対応までは考慮せず，記事全体で類似性を測定している．したがって，本論文の言語横断関連報道記事検索の手法は，既存の研究事例で用いられた手法と比較すると，相対的に簡便な手法であると言える．本論文における評価実験は，言語横断関連報道記事検索の手法として，最も簡便な手法を採用した場合に，どの程度の記事検索性能，および，訳語対応推定性能が達成できるかを示しているということができる．本論文の評価実験において，関連研究で用いられた技術を導入すれば，言語横断関連報道記事検索の性能が向上することが期待できる．具体的には，i) の，言語を横断する際に用いる対訳情報の情報源の種類に関しては，複数の情報源を併用すること，また，ii) に関しては，文レベルの対応まで考慮して記事間の類似度を測定することが考えられる．ただし，本論文の手法は，厳密な文対応付けが困難であるような粗い関連記事群に対しても有効であるという点が長所の一つであると言えるので，ii) の点に関しては，綿密な分析が必要であると思われる．

また，その他の関連研究として，ウェブ上の二言語文書を対象として，URL および HTML 文書の構造における手がかりを利用することにより，対訳で書かれた文書対を収集する手法も提案されている (Resnik and Smith 2003; Nie，Simard，Isabelle and Durand 1999)．

5.2　訳語対応推定

二言語コーパスからの訳語対応推定の手法の研究においては，これまでに，様々な手法が提案されている．本節では，いくつかの観点からそれらの手法を整理するとともに，同一内容の記事組を抽出した後，何らかの形で訳語の対応を推定するという本論文の問題設定に比較的近い研究事例について，本論文の手法との比較を行なう．また，この本論文の問題設定とは独立な観点として，訳語対応を推定する際にどのような情報を用いるかという観点のもとでの整理を行ない，関連研究，および，本論文の手法の間の関係について述べる．

まず，訳語対応推定において用いる要素技術は，大きく分けて，文対応がつけられた対訳コーパスからの訳語対応推定手法，および，コンパラブルコーパスからの訳語対応推定手法という二種類の技術に分けることができる．文対応がつけられた対訳コーパスからの訳語対応推定においては，訳語候補となる語の組に対して，分割表を用いて統計的な相関を測定するとい

う手法がよく知られている (Gale and Church 1991; Kumano and Hirakawa 1994; Haruno and Ikehara 1998; Smadja，McKeown and Hatzivassiloglou 1996; Kitamura and Matsumoto 1996; Melamed 2000)．一方，コンパラブルコーパスからの訳語対応推定においては，一般に，訳語候補となる語の組に対して，何らかの方法で文脈の類似性を測定し，訳語候補の順位付けを行なう．特に，初期の研究 (Fung 1995; Rapp 1995) においては，基本的な語についての既存の対訳辞書を用いずに，文脈の類似性を測定することが試みられたが，以後の研究 (Tanaka and Iwasaki 1996; Fung and Yee 1998; Rapp 1999; 梶，相薗 2001; Chiao and Zweigenbaum 2002; Tanaka 2002; Gaussier，Renders，Matveeva，Goutte and Dejean 2004) では，基本的な語についての既存の対訳辞書を用いて，文脈の類似性を測定している．また，訳語対応推定の研究に関連した研究としては，コンパラブルコーパスを用いて，複数の訳語を持つ語の訳語選択を行なう手法を提案しているものもある (Dagan and Itai 1994; Nakagawa 2001)．

一方，比較的最近の研究においては，要素技術としては，特に，コンパラブルコーパスからの訳語対応推定において用いられた，文脈の類似性に基づく手法を用いるものが多いが，問題設定そのものとして，1) ウェブ上のテキストを利用する，2) 訳語候補の順位付けにおいて，複数の情報源・推定尺度を併用する，といった点に重点を置いた研究事例がいくつかみられる．例えば，ウェブ上のテキストを利用する研究事例としては，ウェブ上で，他言語への翻訳が専門用語に併記されているページを利用して，訳語対応を推定するもの (Nagata，Saito and Suzuki 2001; Cheng，Lu，Teng and Chien 2004) がある．特に，(Cheng et al. 2004) では，英語タームを検索質問として，ウェブ上の中国語ページを収集した結果から中国語訳語候補を生成し，中国語訳語候補と英語タームとの間の統計的相関，および，文脈ベクトルの類似性を併用して，英語・中国語間の訳語対応を推定する手法を提案している．また，(Cao and Li 2002) では，基本語対訳辞書中の訳語の組合せにより，複合語の訳語候補を生成し，ウェブから訳語候補を検索したページから文脈ベクトルを生成して，訳語対応を推定する手法を提案している．また，通常の報道記事をコンパラブルコーパスとして訳語対応を推定する手法の研究においても，英語・中国語間の翻字の情報と文脈ベクトルの類似性を併用して訳語対応を推定するもの (Shao and Ng 2004; Huang，Vogel and Waibel 2004) などがある．

これらの最近の研究の他に，本論文の問題設定に比較的近い研究事例としては，(Fung and Cheung 2004; Munteanu，Fraser and Marcu 2004) がある．これらにおいては，まず，同時期の報道記事をコンパラブルコーパスとして，同一内容の記事組を抽出し，その記事組から対訳文を抽出する．そして，その結果から，統計的機械翻訳モデルを用いて訳語対応を推定する (Fung and Cheung 2004)，あるいは，統計的機械翻訳モデルの性能により，対訳文の質を評価する (Munteanu et al. 2004)，といったことが行なわれる．本論文の問題設定と比較すると，これらの研究事例の問題設定は，同一内容の記事組を抽出した後，何らかの形で訳語の対応を推定するという処理を行なう点が類似していると言える．ただし，最も重要な違いとして，これ

らの研究事例においては，対訳文を抽出する過程を経る必要があるのに対して，本論文の手法においては，記事対応を粗く推定するだけで，訳語対応の推定が可能である点が挙げられる．したがって，本論文の手法は，厳密な文対応付けが困難であるような粗い関連記事群に対しても有効であるという点が特徴であると言える．

また，上で述べた本論文の問題設定とは独立な観点として，訳語対応を推定する際に用いる情報という観点から，関連研究，および，本論文の手法の間の関係を整理することができる．まず，本論文では，訳語対応を推定する際に用いる情報としては,「関連記事組における訳語候補の共起および分割表」(3.1 節) を用いた場合，および,「文脈の類似性」(3.2 節) を用いた場合の比較を行なった．一方，本論文の評価実験で用いなかった情報としては，その他には，複合語の構成要素の訳語対応 (Cao and Li 2002; 吉見，九津見，小谷，佐田，井佐原 2004)，読み等を利用した翻字の情報 (Shao and Ng 2004; Huang et al. 2004; 吉見他 2004) 等がある．さらに，これらの複数の情報を併用することにより，訳語対応推定の性能が改善することが期待できる (日野他 2004b; 吉見他 2004)．

6　おわりに

本論文では，ウェブ上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獲得する手法を提案した．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，関連記事組を用いて二言語間の訳語対応を推定する．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を適用し，評価実験において文脈ベクトルを用いる方法と比較し，この方法が有効であることを示した．本論文では，特に，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．

一方，この評価結果に関連して，特に，報道記事において低頻度であるタームに対しては，訳語対応推定の性能が低下することが分かっている (日野，宇津呂，中川 2004a)．本論文の評価実験において対象としたタームの種類数は高々数百個程度であるが，報道記事に出現するターム全体で言えば，出現頻度が数回程度のタームが相当数あると考えられる．特に，実用的観点から言えば，これらの低頻度タームの訳語をどのようにして獲得するか，という問題を解決することが重要である．この点に関しては，報道記事における出現頻度が小さいタームについては，ウェブ検索エンジンを用いて，ウェブ上での出現文書を収集し，この文書を用いて訳語候補を順位付けることにより，訳語対応推定の性能が改善できることが分かっている (Utsuro, Hino，Kida，Nakagawa and Sato 2004; 木田，宇津呂，日野，佐藤 2004)．また，訳語候補の順位付けにおいて，正解訳語を必ずしも一位に順位付けすることができなくても，上位 10 位以

内程度に正解訳語を含めることができれば，人間が訳語対応を発見する過程を比較的効率的に支援できると考えられる．この点に関しては，訳語対応推定の結果，および，各タームが出現する文書を閲覧する機能を備えた訳語対応獲得支援インタフェース (Utsuro，Horiuchi，Chiba and Hamamoto 2002; 日野，堀内，浜本，中山，宇津呂 2003) を援用できると考えている．本論文で提案した技術を利用する局面としては，直接的には，機械翻訳用の対訳辞書を強化することが挙げられるが，その他に，例えば，言語横断情報検索において，既存の対訳辞書や機械翻訳システムに未登録の訳語組を収集する場合や，人間の翻訳者が必要とする翻訳知識を収集する場合などが考えられる．

参考文献

Cao, Y. and Li, H. (2002). “Base Noun Phrase Translation Using Web Data and the EM Algorithm.” In *Proceedings of the 19th COLING*, pp. 127–133.

Cheng, P.-J., Lu, W.-H., Teng, J.-W., and Chien, L.-F. (2004). “Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora.” In *Proceedings of the 42nd ACL*, pp. 534–541.

Chiao, Y.-C. and Zweigenbaum, P. (2002). “Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora.” In *Proceedings of the 19th COLING*, pp. 1208–1212.

Collier, N., Hirakawa, H., and Kumano, A. (1998). “Machine Translation vs. Dictionary Term Translation — A Comparison for English-Japanese News Article Alignment.” In *Proceedings of the 17th COLING and the 36th Annual Meeting of ACL*, pp. 263–267.

Dagan, I. and Itai, A. (1994). “Word Sense Disambiguation Using a Second Language Monolingual Corpus.” *Computational Linguistics*, **20** (4), 563–596.

Fung, P. (1995). “Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus.” In *Proceedings of 3rd Workshop on Very Large Corpora*, pp. 173–183.

Fung, P. and Cheung, P. (2004). “Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM.” In *Proceedings of EMNLP*, pp. 57–63.

Fung, P. and Yee, L. Y. (1998). “An IR Approach for Translating New Words from Nonparallel, Comparable Texts.” In *Proceedings of the 17th COLING and the 36th Annual Meeting of ACL*, pp. 414–420.

Gale, W. and Church, K. (1991). “Identifying Word Correspondences in Parallel Texts.” In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp. 152–157.

Gaussier, E., Renders, J., Matveeva, I., Goutte, C., and Dejean, H. (2004). “A Geometric

View on Bilingual Lexicon Extraction from Comparable Corpora." In *Proceedings of the 42nd ACL*, pp. 526–533.

浜本武，中山健明，日野浩平，堀内貴司，宇津呂武仁 (2003). "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.

Haruno, M. and Ikehara, S. (1998). "Two-Step Extraction of Bilingual Collocations by Using Word-Level Sorting." 電子情報通信学会論文誌, **E81-D** (10), 1103–1110.

Hasan, M. M. and Matsumoto, Y. (2001). "Multilingual Document Alignment — A Study with Chinese and Japanese." In *Proceedings of the 6th NLPRS*, pp. 617–623.

日野浩平，堀内貴司，浜本武，中山健明，宇津呂武仁 (2003). "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, pp. 421–424.

日野浩平，宇津呂武仁，中川聖一 (2004a). "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." 情報処理学会研究報告, **2004** ((2004–NL–162)), 57–63.

日野浩平，宇津呂武仁，中川聖一 (2004b). "日英報道記事からの訳語対応推定における複数の推定尺度の利用." 言語処理学会第 10 回年次大会論文集, pp. 249–252.

堀内貴司，千葉靖伸，浜本武，宇津呂武仁 (2002a). "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** ((2002–NL–150)), 191–198.

堀内貴司，千葉靖伸，浜本武，宇津呂武仁 (2002b). "翻訳知識獲得のための言語横断関連報道記事検索." 言語処理学会第 8 回年次大会論文集, pp. 303–306.

堀内貴司，日野浩平，浜本武，中山健明，宇津呂武仁 (2003). "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." 言語処理学会第 9 回年次大会論文集, pp. 341–344.

Huang, F., Vogel, S., and Waibel, A. (2004). "Improving Named Entity Translation Combining Phonetic and Semantic Similarities." In *Proceedings of HLT-NAACL*, pp. 281–288.

梶博行，相薗敏子 (2001). "共起集合の類似度に基づく対訳コーパスからの対訳語抽出." 情報処理学会論文誌, **42** (9), 2248–2258.

木田充洋，宇津呂武仁，日野浩平，佐藤理史 (2004). "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, **2004** ((2004–NL–162)), 65–70.

Kitamura, M. and Matsumoto, Y. (1996). "Automatic Extraction of Word Sequence Correspondences in Parallel Corpora." In *Proceedings of the 4th Workshop on Very Large Corpora*, pp. 79–87.

Kumano, A. and Hirakawa, H. (1994). "Building an MT Dictionary from Parallel Texts based on Linguistic and Statistical Information." In *Proceedings of the 15th COLING*, pp. 76–81.

Masuichi, H., Flournoy, R., Kaufmann, S., and Peters, S. (2000). "A Bootstrapping Method for Extracting Bilingual Text Pairs." In *Proceedings of the 18th COLING*, pp. 1066–1070.

Matsumoto, K. and Tanaka, H. (2002). "Automatic Alignment of Japanese and English Newspaper Articles using an MT System and a Bilingual Company Name Dictionary." In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Vol. 2, pp. 480–484.

Matsumoto, Y. and Utsuro, T. (2000). "Lexical Knowledge Acquisition." In Dale, R., Moisl, H., and Somers, H. (Eds.), Handbook of Natural Language Processing, chap. 24, pp. 563–610. Marcel Dekker Inc.

Melamed, I. D. (2000). "Models of Translational Equivalence among Words." *Computational Linguistics*, **26** (2), 221–249.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). "Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora." In *Proceedings of HLT-NAACL*, pp. 265–272.

Nagata, M., Saito, T., and Suzuki, K. (2001). "Using the Web as a Bilingual Dictionary." In *Proceedings of the ACL-2001 Workshop on Data-driven Methods in Machine Translation*, pp. 95–102.

Nakagawa, H. (2001). "Disambiguation of Single Noun Translations Extracted from Bilingual Comparable Corpora." *Terminology*, **7** (1), 63–83.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). "Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web." In *Proceedings of the 22nd SIGIR*, pp. 74–81.

Pei, J., Han, J., Mortazavi-Asl, B., and Pinto, H. (2001). "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth." In *Proceedings of the 17th International Conference on Data Mining*, pp. 215–224.

Rapp, R. (1995). "Identifying Word Translations in Non-Parallel Texts." In *Proceedings of the 33rd Annual Meeting of ACL*, pp. 320–322.

Rapp, R. (1999). "Automatic Identification of Word Translations from Unrelated English and German Corpora." In *Proceedings of the 37th Annual Meeting of ACL*, pp. 519–526.

Resnik, P. and Smith, N. (2003). "The Web as a Parallel Corpus." *Computational Linguistics*, **29** (3), 349–380.

Shao, L. and Ng, H. T. (2004). "Mining New Word Translations from Comparable Corpora." In *Proceedings of the 20th COLING*, pp. 618–624.

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

Takahashi, Y., Shirai, S., and Bond, F. (1997). "A Method of Automatically Aligning Japanese and English Newspaper Articles." In *Proceedings of the 4th NLPRS*, pp. 657–660.

Tanaka, K. and Iwasaki, H. (1996). "Extraction of Lexical Translations from Non-Aligned Corpora." In *Proceedings of the 16th COLING*, pp. 580–585.

Tanaka, T. (2002). "Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora." In *Proceedings of the 19th COLING*, pp. 981–987.

内山将夫，井佐原均 (2003). "日英新聞の記事および文を対応付けるための高信頼性尺度." 自然言語処理, **10** (4), 201–220.

Utsuro, T., Horiuchi, T., Chiba, Y., and Hamamoto, T. (2002). "Semi-automatic Compilation of Bilingual Lexicon Entries from Cross-Lingually Relevant News Articles on WWW News Sites." In Richardson, S. D. (Ed.), *Machine Translation: From Research to Real Users*, Lecture Notes in Artificial Intelligence: Vol. 2499, pp. 165–176. Springer.

Utsuro, T., Horiuchi, T., Hamamoto, T., Hino, K., and Nakayama, T. (2003). "Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora." In *Proceedings of the 10th EACL*, pp. 355–362.

Utsuro, T., Hino, K., Kida, M., Nakagawa, S., and Sato, S. (2004). "Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition." In *Proceedings of the 20th COLING*, pp. 1036–1042.

Xu, D. and Tan, C. L. (1999). "Alignment and Matching of Bilingual English-Chinese News Texts." *Machine Translation*, **14**, 1–33.

吉見毅彦，九津見毅，小谷克則，佐田いち子，井佐原均 (2004). "複合語の内部情報・外部情報を統合的に利用した訳語対の抽出." 自然言語処理, **11** (4), 89–103.

略歴

宇津呂 武仁: 1989 年京都大学工学部 電気工学第二学科 卒業．1994 年同大学大学院工学研究科 博士課程電気工学第二専攻 修了．京都大学博士 (工学)．奈良先端科学技術大学院大学情報科学研究科助手，豊橋技術科学大学工学部情報工学系講師を経て，2003 年より 京都大学 情報学研究科 知能情報学専攻講師．自然言語処理の研究に従事．

日野 浩平: 2003 年豊橋技術科学大学 工学部 情報工学系卒業．2005 年 同大学

大学院工学研究科修士課程 情報工学専攻修了．現在，NTT データテクノロジー株式会社に勤務．在学中は自然言語処理に関する研究に従事．

堀内 貴司: 2001 年豊橋技術科学大学 工学部 情報工学系卒業．2003 年 同大学大学院工学研究科修士課程 情報工学専攻修了．現在，日立製作所に勤務．在学中は自然言語処理に関する研究に従事．

中川 聖一: 1976 年 京都大学大学院工学研究科博士課程修了．同年京都大学工学部 情報工学科 助手．1980 年 豊橋技術科学大学工学部情報工学系講師．1990 年 同教授．1985∼1986 年 カーネギーメロン大学客員研究員．音声言語情報処理，自然言語処理，人工知能の研究に従事．工学博士．1977 年 電子通信学会論文賞，1998 年度 IETE 最優秀論文賞，2001 年 電子情報通信学会論文賞受賞．著書「確率モデルによる音声認識」（電子情報通信学会編）,「音声・聴覚と神経回路網モデル」（共著，オーム社）,「情報理論の基礎と応用」（近代科学社）,「パターン情報処理 」（丸善）など．

(2004 年 12 月 2 日 受付)
(2005 年 2 月 14 日 再受付)
(2005 年 5 月 5 日 再々受付)
(2005 年 5 月 28 日 採録)