自然言語処理　150－28
(2002. 7. 16)

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

堀内　貴司† 千葉　靖伸†† 浜本　武† 宇津呂武仁†

† 豊橋技術科学大学 工学部 情報工学系
〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
†† アライドテレシス株式会社
E-mail: †,††{takashi,chiba,hamamo,utsuro}@cl.ics.tut.ac.jp

あらまし 本論文では，WWW 上の報道記事サイト等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・言い回しなどの訳語対応を半自動的に獲得する枠組を提案する．特に本論文では，言語を横断して内容的に関連した日英報道記事を収集する手法について述べ，さらに，言語横断関連報道記事検索により自動収集された日英関連記事対から，半自動的に訳語対応を獲得する手法を提案する．評価実験においては，評価用記事集合に対して言語横断関連報道記事検索の性能を評価した後，言語横断関連報道記事検索の性能と訳語対応獲得の性能の相関について分析した結果について詳しく述べる．

キーワード 機械翻訳，訳語対応獲得，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

Acquisition of Bilingual Term Correspondences from Relevant Japanese-English News Articles Automatically Collected by CLIR

Takashi HORIUCHI†, Yasunobu CHIBA††,

Takeshi HAMAMOTO†, and Takehito UTSURO†

† Department of Information and Computer Sciences,
Faculty of Engineering, Toyohashi University of Technology
Tenpaku-cho, Toyohashi, Aichi 441–8580, Japan
†† Allied Telesis K.K.
E-mail: †,††{takashi,chiba,hamamo,utsuro}@cl.ics.tut.ac.jp

Abstract For the purpose of overcoming resource scarcity bottleneck in corpus-based translation knowledge acquisition research, this paper takes an approach of semi-automatically acquiring domain specific translation knowledge from the collection of bilingual news articles on WWW news sites. This paper presents results of applying standard co-occurrence frequency based techniques of estimating bilingual term correspondences to relevant article pairs automatically collected from WWW news sites. The experimental evaluation results are very encouraging and it is proved that many useful bilingual term correspondences can be efficiently discovered with little human intervention from relevant article pairs on WWW news sites.

Key words machine translation, acquisition of bilingual term correspondences, parallel corpus, comparable corpus, cross-language IR, bilingual lexicon

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．

本研究における WWW からの翻訳知識獲得の流れを図 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 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．本論文では特に，報道内容がほぼ同一の日英記事対のことを “同一内容” の二言語記事とよび，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” の二言語記事とよぶ．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，その推定範囲から翻訳知識を獲得する (ただし，本論文の評価実験の範囲では，内容的に対応する翻訳部分の推定は行っていない)．

この一連の枠組において，特に本論文では，WWW 上の新聞社やテレビ局のサイトから日本語および英語で書かれた報道記事を取得し，言語を横断して内容的に関連した日英報道記事を収集する手法について述べる．さらに，言語横断関連報道記事検索により自動収集された日英関連記事対から，半自動的に訳語対応を獲得する手法について述べる．特に，訳語対応の半自動獲得においては，作業者が，構造化された訳語対応推定結果を走査しながら，必要に応じて，言語横断関連報道記事検索により自動収集された日英関連記事対を閲覧することにより，正しい訳語対応を効率よく選び出すという枠組を提案する．また，評価実験においては，評価用記事集合に対して言語横断関連報道記事検索の性能を評価した後，言語横断関連報道記事検索の性能と訳語対応獲得の性能の相関について分析した結果について詳しく述べる．

2. 言語横断関連報道記事検索

日英関連記事対検索の流れを図 2 に示す．まず，新聞社やテレビ局のサイトから英語記事と日本語記事を取得する．そして，市販の翻訳ソフト(注1)を用いて英語記事を日本語に翻訳する．次に，翻訳ソフトにより日本語訳した記事と取得してきた日本語記事を，日本語形態素解析システム「茶筌」[6] によって形態素解析し，形態素の頻度ベクトルを作成する．そして，頻度ベクトル間で余弦類似度を計算し(注2)，類似度が上位の記事対を検索結果とする．その際，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(注3)．なお，翻訳ソフトとして，日英翻訳ソフトを用いて日本語記事を英語に翻訳し，英語単語の頻度ベクトル間で類似度計算を行うことも可能であるが，本論文では，手近に利用可能な翻訳ソフトのうち，バッチ処理機能付きのものの種類は，英日翻訳ソフトの方が多かったので，英日翻訳ソフトを用いて英語記事を日本語に翻訳する方式をとった．

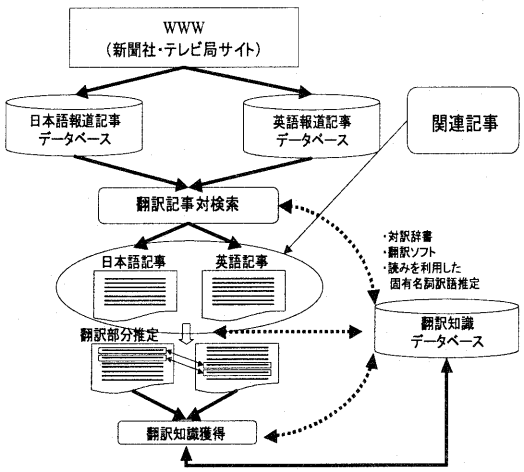


図 1: WWW からの翻訳知識獲得の流れ

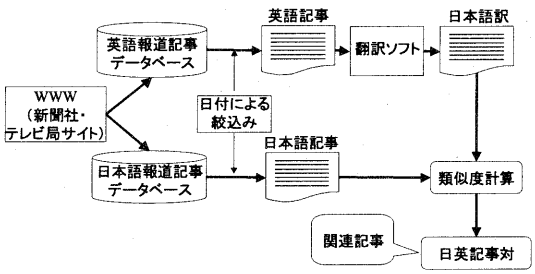


図 2: 日英関連記事対検索の流れ

3. 日英関連報道記事からの訳語対応の獲得

3.1 訳語対応の推定

本節では，2. 節で述べた言語横断関連報道記事検索により自動収集された日英関連記事対から，訳語対応を推定する手法について述べる．4.1 節で述べるように，日本国内の新聞社・テレビ局等の報道サイトでは，一日に掲載される記事数は日本語記事の方が英語記事よりも約 5~30 倍ほど多い．したがって，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，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を比べると，前者の方がはるかに収集効率がよい．このことをふまえて，本節でも，英語記事を検索質問として関連日本語記事を収

(注1)：市販の翻訳ソフトとしては，バッチ処理機能付きのものを数種類比較したが，言語横断関連記事検索における性能に大きな差はなかった．その中で，オムロンソフトウェア社製「翻訳魂」の性能が，他の性能を若干上回っていたため，本論文の評価実験においては同翻訳ソフトを用いた．

(注2)：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．

(注3)：複数日掲載記事については，初掲載の日付だけを掲載日とした．

集した結果から訳語対応を推定するという方法をとる．

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．

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

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 PPC_{EJ} とする．

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

以下では，この集合 PPC_{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．

本論文では特に，この疑似的対訳コーパス PPC_{EJ} に対して，文献 [5] で紹介されているような，共起頻度を用いた標準的訳語対応推定尺度を適用し，その有効性を評価する．まず，英語および日本語の単言語での連語 (もしくは単語)t_E および t_J に対して，頻度の下限，および，連語の場合の構成単語数の上限を設定し，これらの条件を満たす単言語での連語もしくは単語を抽出する．そして，以下の 2×2 分割表を用いて，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 [5] を用いて訳語対応推定を行う．

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

実験において上記の四種類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，対数尤度比が最もよく，dice 係数，相互情報量という順で性能が劣化した．そこで，4.3 節では，特に，以下の ϕ^2 統計を用いて訳語対応を推定した結果について分析する．

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

3.2 訳語対応の半自動獲得

次に，本節では，前節で述べた手法により疑似的対訳コーパス PPC_{EJ} から訳語対応を推定した結果から，適切な訳語対応を半自動的に獲得する手順の概要を述べる．本論文では，自動収集された日英関連報道記事対を疑似的対訳コーパスとみなして，通常の対訳コーパスからの訳語対応推定手法を適用するが，疑似的対訳コーパスには内容的に無関係な記事も多く含まれるため，全自動で質の高い訳語対応を獲得することは難しい．そこで，本論文では，訳語対応推定結果の中から，正しい訳語対応を効率よく選び出すことを実現するために，以下の二つの基準に基づいて，推定された訳語対応全体を部分集合に分割して構造化した上で訳語対応推定結果を走査する．

(1) 訳語対応推定結果の訳語組のうち，英語側の連語または単語 (あるいは日本語側の連語または単語) が共通の訳語組をまとめる．

(2) 英語側の連語 (あるいは日本語側の連語) の間の包含関係を考慮し，ある連語または単語が別の連語の一部になっているという包含関係が成り立つ場合には，それらの連語または単語に関する訳語対応推定結果をまとめる．

具体的には，まず，ある連語または単語 t が別の連語または単語 t' と同一であるか，または，その一部を構成するという関係を $t \succeq t'$ で記述する(注4)．そして，ある英語の連語もしくは単語 t_E について，他のどの英語の連語 $t'_E (\neq t_E)$ に対しても，その一部を構成しない ($t \not\succeq t'$) 場合に，以下の手順で訳語組の集合 $TP(t_E)$ を構成し，訳語対応推定結果全体の集合を (互いに素とは限らない) 部分集合に分割する．

$$TP(t_E) = \left\{ \langle t'_E, t_J \rangle \mid t_E \succeq t'_E, freq(t_E) \geqq L_f^E, freq(t_J) \geqq L_f^J, \right.$$
$$\left. freq(t_E, t_J) \geqq L_f^{EJ}, length(t_E) \leqq U_l^E, length(t_J) \leqq U_l^J \right\}$$

ここで，L_f^E，L_f^J，L^{EJ_f} はそれぞれ，英語連語または単語の頻度下限，日本語連語または単語の頻度下限，日英間の共起頻度下限，また，U_l^E および U_l^J はそれぞれ，英語連語および日本語連語を構成する単語数の上限である．集合 $TP(t_E)$ は，英語の連語もしくは単語 t_E に対して，t_E もしくはその一部を構成する語が英語側の語となっている訳語対応推定結果のうち，頻度下限および構成単語数の上限を満たす訳語組を集めた集合である．このとき，語 t_E をインデックス語とよぶ (ここでは，インデックス語が英語の場合について説明するが，インデックス語が日本語の場合でも同様の手順で訳語対応推定結果の評価を行うことができる)．

次に，各集合 $TP(t_E)$ に対して，要素となっている訳語組の ϕ^2 統計の値のうちの最大値を $\hat{\phi^2}(TP(t_E))$ とする．

$$\hat{\phi^2}(TP(t_E)) = \max_{\langle t_E, t_J \rangle \in TP(t_E)} \phi^2(t_E, t_J)$$

そして，全ての集合 $TP(t_E^1), \ldots, TP(t_E^m)$ を $\hat{\phi^2}(TP(t_E))$ の値の大きい順に並べ，先頭から順に各集合 $TP(t_E)$ の要素を人手で調べていき，各集合 $TP(t_E)$ が正しい訳語対を含む率を評価することとする．

$$\begin{matrix}\text{正しい} \\ \text{訳語対を} \\ \text{含む率}\end{matrix} = \frac{|\{TP(t_E) \mid \text{正しい訳語対 } \langle t_E, t_J \rangle \in TP(t_E)\}|}{|\{TP(t_E) \mid TP(t_E) \neq \emptyset\}|} \quad (1)$$

また，各集合 $TP(t_E)$ の要素である訳語対を人手で調べる際には，ϕ^2 統計の値の大きい順に調べることとし，ϕ^2 統計の値の大小と訳語対応の正誤の相関についても評価する．

3.3 例

前節の手順により構造化された訳語対応推定結果を走査しながら，必要に応じて，言語横断関連報道記事検索により自動収集された日英関連記事対を閲覧することにより，正しい訳語対応を効率よく選び出す様子を図 3 に示す．まず，図 3 の上部には，インデックス語 t_E が英語の連語 “Tokyo District Court” の場合について，ϕ^2 統計値の大きい順に，日本語側の語 t_J，頻度 $freq(t_E)$，$freq(t_J)$，$freq(t_E, t_J)$，および，$\phi^2(t_E, t_J)$ の値を示す．この例の場合には，$\phi^2(t_E, t_J)$ の値が最も大きい「東京地裁」がインデックス語 t_E の正しい訳語となっている．作業者は，必要に応じて，任意の訳語推定結果の組 t_E と t_J を選び，t_E および t_J をそれぞれ含み，余弦類似度の下限の条件

(注4)：日本語の単語の単位は，日本語形態素解析システム「茶筌」[6] の形態素の単位とする．

表 1: 記事の日数・記事数・平均記事長・評価用記事対数

サイト	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	
	英語	日本語	英語	日本語	英語	日本語	英語	日本語	同一内容	関連話題
A	562	578	607	21349	1.1	36.9	1087.3	759.9	24	33
B	162	168	2910	14854	18.0	88.4	3135.5	836.4	28	82
C	162	166	3435	16166	21.2	97.4	3228.9	837.7	28	31

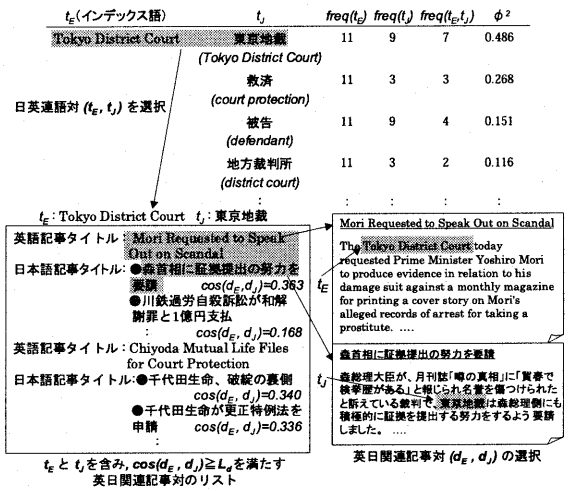


図 3: 日英関連報道記事対からの訳語対応の半自動獲得の例

$\cos(d_E, d_J) \geq L_d$ を満たす記事組 d_E と d_J を閲覧することにより，t_E と t_J が訳語組として適切であるか否かを判断することができる．その際には，図 3 左下に示すように，インデックス語 t_E を含む英語記事のタイトルのリストが表示され，各英語記事に対して，余弦類似度の下限の条件を満たし，かつ日本語の語 t_J を含む日本語記事のタイトルのリストが表示される．作業者は，これらの日英関連記事中での語 t_E および t_J の使われ方を閲覧することにより，t_E と t_J が適切な訳語対であるか否かを効率よく判断することができる．また，t_E と t_J が適切な訳語対でない場合でも，選択した記事組 d_E と d_J が同一または関連した内容の報道記事であれば，容易に適切な訳語対を発見することができる．もし，選択した記事組 d_E と d_J の内容があまり関連していない場合には，より適切な記事組を選択することにより，訳語対発見の作業を継続することが可能である．

4. 評　　価

4.1 日英関連報道記事対の収集

本節では，評価実験で用いた日英報道記事について述べる．本論文の評価実験では，A～C の三種類のサイトから収集した日本語および英語の報道記事を用いた．まず，各サイトにおいて収集対象となった記事の総日数，総記事数，一日の平均記事数，および一記事の平均記事長を表 1 に示す．総記事数，一日の平均記事数については，3 サイトとも英語記事よりも日本語記事の方が多い．次に，言語横断関連報道記事検索の性能評価 (詳細は次節参照) のために人手で収集した評価用日英関連記事対

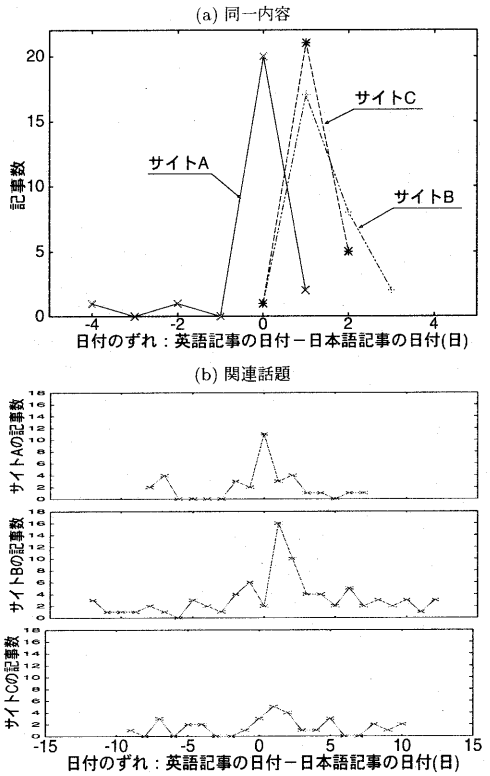


図 4: 日英関連記事間の日付のずれの分布

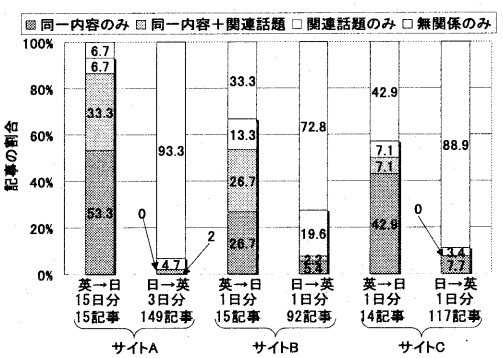


図 5: 相手言語における同一内容・関連話題・無関係記事の有無の割合

の数の内訳を表 1 の「評価用記事対数」の欄に示す(注5)．また，

(注5)：人手による日英関連記事対の収集においては，文献 [3] で紹介した言語横断関連報道記事検索・収集・閲覧システムを用いた．ただし，英語記事の方が

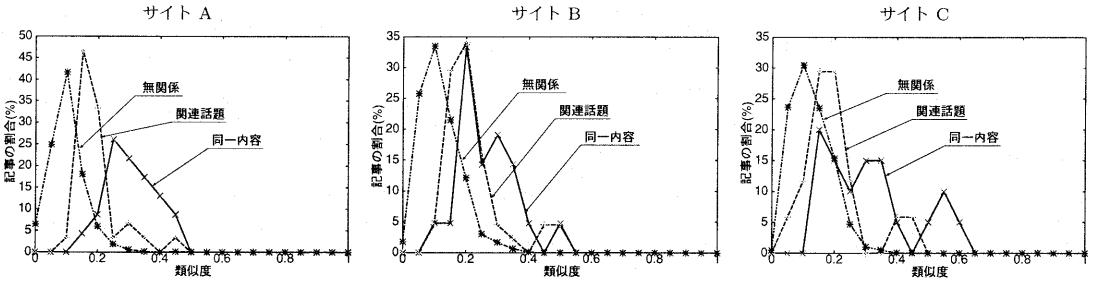


図 6: 記事間類似度の分布 (英語 → 日本語)

評価用関連記事対の日英記事間の日付のずれの分布を図 4 に示す．この結果から分かるように,「同一内容」の記事間の日付のずれは ± 数日であるのに対して,「関連話題」の記事間の日付のずれは ±10 日前後に及ぶ．

次に，一つの記事に対して，相手言語側に「同一内容」あるいは「関連話題」の記事が実際に存在する割合を調査した結果を図 5 に示す．この調査においては，適当な日数の範囲で (初) 掲載された全記事 (サイト A: 英日検索では 15 日分 15 記事，日英検索では 3 日分 149 記事，サイト B: 英日検索では 1 日分 15 記事，日英検索では 1 日分 92 記事，サイト C: 英日検索では 1 日分 14 記事，日英検索では 1 日分 117 記事) に対して，図 4 に示した最大日付幅の範囲において言語横断関連報道記事検索を行い，類似度上位 40 以内に，

i) 「同一内容」記事が少なくとも一つ存在し
「関連話題」記事が一つも存在しない記事数
ii) 「同一内容」記事および「関連話題」記事が
それぞれ少なくとも一つ存在する記事数
iii) 「関連話題」記事が少なくとも一つ存在し
「同一内容」記事が一つも存在しない記事数
iv) 「同一内容」記事または「関連話題」記事が
一つも存在しない記事数

をそれぞれ集計してその分布を求めた．この結果から分かるように，いずれのサイトにおいても，英語記事よりも日本語記事の方がその数が多いために，日英検索において何らかの関連記事が存在する割合は，10〜30%前後と低くなっているのに対して，英日検索においては，半数以上の英語記事に対して「同一内容」の記事が日本語側に存在し,「関連話題」の記事を含めると，その割合は 10%弱 ～ 数 10%程度増える．この結果から，英語記事から日本語記事を検索する方向で言語横断関連報道記事収集を行えば，5 割以上の率で有用な日英記事対が収集できることが分かる．さらに，図 5 において調査対象となった全記事対に対して,「同一内容」記事対,「関連話題」記事対,「無関係」記事対の各々について，記事間類似度の分布をプロットした結果を図 6 に示す．三つのグループの間で類似度の平均値には一定の差があることが伺えるものの，分布の重複も一定量存在している．

記事数が少なく，英語記事を検索質問記事とした方が関連する日本語記事が存在する確率が高いため，検索質問記事は英語記事とした．日英関連記事対の収集効率は，一時間あたりの収集記事対数が 15 程度という，十分高いものであった．

4.2 言語横断関連報道記事検索

表 1 の評価用記事対に対して，言語横断関連報道記事検索の性能の評価を行った．評価用記事対の英語記事を検索質問として，記事対の日本語記事を含む記事集合に対して言語を横断した記事検索を行い，上位 n 位以内に関連記事が含まれる率 (上位 n 位以内の再現率) を測定し，順位 n に対する再現率の変化をプロットした結果を図 7 に示す(注6)．この際，検索対象記事の日付の範囲については，図 4 に示した最大日付幅の場合 (「同一内容」: 日付のずれ ±4 日 (サイト A)，±3 日 (サイト B)，±2 日 (サイト C),「関連話題」: 日付のずれ ±8 日 (サイト A)，±12 日 (サイト B)，±10 日 (サイト C))，および，日付幅をある程度絞り込んだ場合 (「同一内容」: 日付のずれ ±1 日 (サイト A，B，C),「関連話題」: 日付のずれ ±4 日 (サイト A)，±6 日 (サイト B)，±5 日 (サイト C)) の二通りの結果を示す．日付の範囲を絞り込んだ場合の再現率の定義は，日付の範囲内の評価用 (同一内容または関連話題) 記事対の集合を DP_{ref} とすると，

$$\text{再現率} = \frac{\left|\left\{\langle d_E, d_J\rangle \mid \langle d_E, d_J\rangle \in DP_{ref},\ \text{検索質問 } d_E \text{に対して } d_J \text{が上位 } n \text{ 位内に含まれる}\right\}\right|}{|DP_{ref}|}$$

となる．評価結果においては,「同一内容」「関連話題」のいずれにおいても，検索対象記事の日付の範囲が小さい方が，誤検索となる「無関係」記事数が少ないために，検索性能はよい．図 7 の結果において，日付幅最大での「同一内容」記事検索の性能は，サイト A および C では上位 20 位以内で再現率 100%であり，サイト B においても上位 40 位以内で再現率 90%以上である．

一方，図 8 には，図 4 に示した最大日付幅の場合について，記事間の類似度の下限値の条件を満たす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．この場合の適合率・再現率の定義は，記事間類似度の下限値を L_d として，

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

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

となる．図 8 の結果から,「同一内容」の記事の場合，サイト

(注6)：日本語記事を検索質問として英語関連記事を検索する場合には，誤検索となる「無関係」記事数がより少ないため，英語記事を検索質問とする場合よりも検索性能はよい [3].

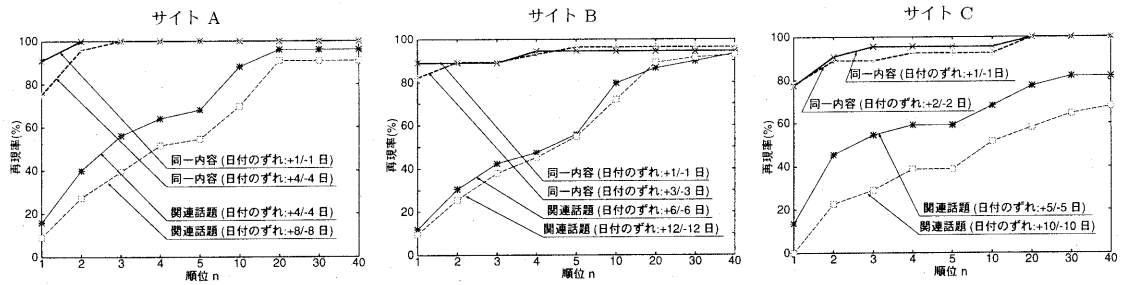


図 7: 日英関連記事検索の再現率 (上位 n 位内)

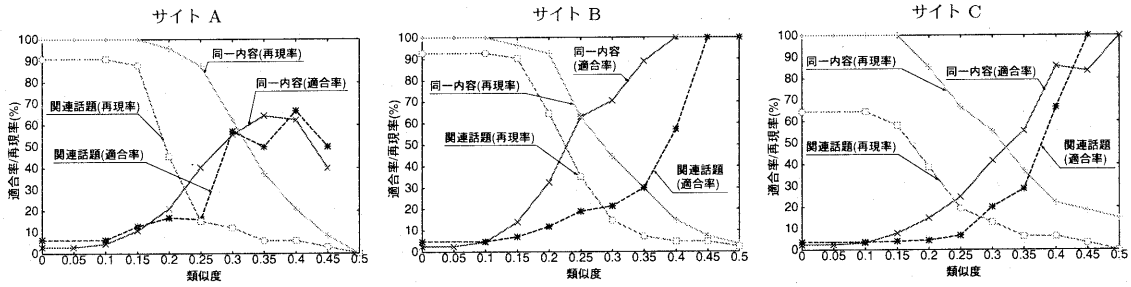


図 8: 日英関連記事検索の適合率・再現率 (記事間類似度 $\geq L_d$, 図 4 の最大日付幅)

表 2: 記事間類似度の下限の条件を満たす日英報道記事の数

サイト	A				B		C	
類似度下限 L_d	0.25	0.3	0.4	0.5	0.4	0.5	0.4	0.5
日付のずれ (日)	± 4				± 3		± 2	
英語記事数	473	362	190	74	415	92	453	144
日本語記事数	1990	1128	377	101	631	127	725	185

A と比較すると，サイト B，C では，類似度の下限値が 0.4 以上では，80%以上の高い適合率が達成できていることが分かる．したがって，サイト B，C では，類似度の下限値が高い場合に，無関係記事がほとんど混入しないという条件で訳語対応の推定が行えることが期待できる．

4.3 日英関連報道記事からの訳語対応の半自動獲得

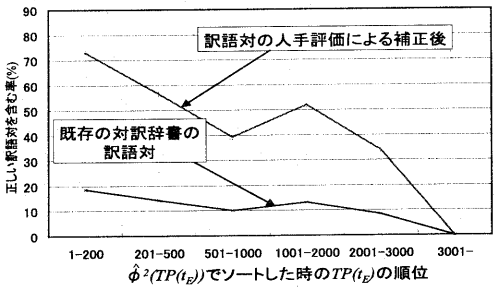
サイト A，B，C の日英報道記事に対して，記事間類似度の下限値 L_d のいくつかの設定のもとで，英語記事を検索質問とした言語横断関連報道記事検索により関連日本語記事を自動収集し，その結果から，3.2 節の手順に基づいて訳語対応の半自動獲得を行う過程を評価した．まず，表 2 に，記事間の日付のずれを図 4(a) に示した「同一内容」記事対の最大日付幅とした場合に，記事間類似度の下限値 L_d の条件を満たす日英報道記事の数を示す (一つの日本語記事が二つ以上の英語記事により検索される場合には，記事数を重複して数える)．そして，頻度下限値および連語構成単語数の上限値として $L_f^E = L_f^J = 3, L_f^{EJ} = 2, U_l^E = U_l^J = 5$ という条件で，訳語組の集合 $TP(t_E)$ を構成し，式 (1) の「集合 $TP(t_E)$ が正しい訳語対を含む率」を評価した．

まず，図 9 に，英語の語 t_E をインデックス語として訳語対の集合 $TP(t_E)$ を構成した場合について，$\hat{\phi}^2(TP(t_E))$ の値の大きい順に集合 $TP(t_E)$ をソートした順位に対する，「集合 $TP(t_E)$ が正しい訳語対を含む率」の分布を示す．図 9(a) には，サイト A，記事間類似度下限値 $L_d = 0.3$ の場合に，既存の対訳辞書 (英辞郎 Ver.37: 85 万語) の訳語対が集合 $TP(t_E)$ に含まれる率の分布を示す．また，$\hat{\phi}^2(TP(t_E))$ の上位 200 位以内の集合 $TP(t_E)$ を人手で評価して(注7)，正しい訳語対を含む率を人手で算出した．さらに，この人手評価による「正しい訳語対を含む率」と，「既存の対訳辞書の訳語対の含有率」との比率を，「既存の対訳辞書の訳語対の含有率」に一律にかけることにより，人手評価による「正しい訳語対を含む率」の推定値を算出し，その分布を示す (以降の図 9(b)，(c)，および図 10 では，人手評価による「正しい訳語対を含む率」の推定値のみをプロットしているが，その推定値は，いずれも，ここで用いた比率を既存の対訳辞書の訳語対の含有率にかけて算出したものである)．この結果より，既存の対訳辞書に含まれる訳語対の約 2.5 倍の訳語対が含まれていることがわかる．また，「正しい訳語対を含む率」は，$\hat{\phi}^2(TP(t_E))$ の順位が高い方が大きい傾向にあり，ϕ^2 統計値の有効性が確認できる．

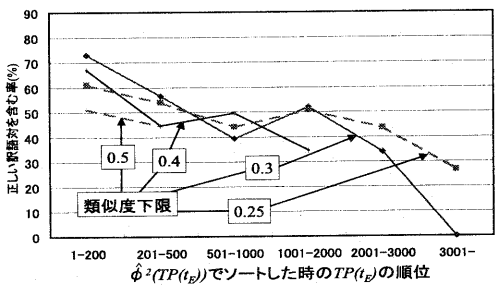
次に，サイト A において記事間類似度の下限値 L_d を 0.25，0.3，0.4，0.5 と変化させた場合の比較，および，サイト B，C において，記事間類似度の下限値 L_d を 0.4，0.5 と変化させた場合の比較を，英語の語 t_E がインデックス語の場合 (図 9(b)，(c))，および，日本語の語 t_J がインデックス語の場合 (図 10(a)，(b)) の両方について示す．これらの結果では，記事間類似度の下限値が小さくなるほど対象とする記事数が増えるため，イン

(注7)：この評価結果では，訳語対の連語長の内訳は，日本語の方が，1 語: 18%，2 語: 59%，3 語: 15%，4 語: 7%，5 語: 1%で，英語の方が，1 語: 26%，2 語: 65%，3 語: 8%，4 語: 1%であった．また，訳語対の約 35%が人名，組織名などの固有表現であった．

(a) サイト A，記事間類似度下限 $L_d = 0.3$，
既存の対訳辞書の訳語対/
$\hat{\phi}^2(TP(t_E))$ 上位 200 個の $TP(t_E)$ の訳語対の人手評価による補正後



(b) サイト A，記事間類似度下限 $L_d = 0.25, 0.3, 0.4, 0.5$



(c) サイト B・C，記事間類似度下限 $L_d = 0.4, 0.5$

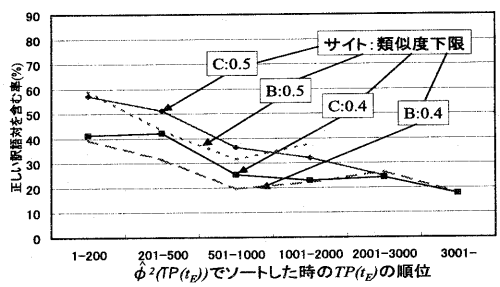
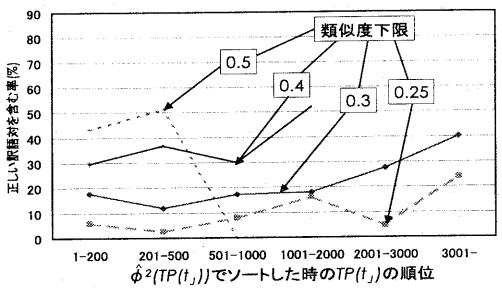


図 9: 訳語対の集合 $TP(t_E)$ に正しい訳語対を含む率の分布 (インデックス語: 英語)

デックス語の数が多くなり，プロットが長くなっている．また，インデックス語が英語の場合と日本語の場合を比較すると，インデックス語が英語の場合は，類似度下限値が小さくなっても，$\hat{\phi}^2(TP(t_E))$ の順位が高い方が「正しい訳語対を含む率」が大きいという傾向を保持するのに対して，インデックス語が日本語の場合は，類似度下限値が小さくなるほど，$\hat{\phi}^2(TP(t_J))$ の順位が高い場合の「正しい訳語対を含む率」が大きく劣化するという傾向がある．これは，表 2 から分かるように，類似度下限値が小さくなるほど，英語記事とは無関係な内容の日本語記事の比率が大きくなり，それに伴って，英語記事に適切な訳語を持たない日本語のインデックス語が増大することが原因である．また，サイト B とサイト C の比較では，特にインデックス語が日本語，記事間類似度の下限値 $L_d = 0.5$ の場合で，$\hat{\phi}^2(TP(t_J))$ の順位が高い部分で，サイト C がサイト B を引き離している．これは，表 2 から分かるように，この条件では，サイト C の方が記事数が多く，かつ，英語記事とは無関係な内

(a) サイト A，記事間類似度下限 $L_d = 0.25, 0.3, 0.4, 0.5$



(b) サイト B・C，記事間類似度下限 $L_d = 0.4, 0.5$

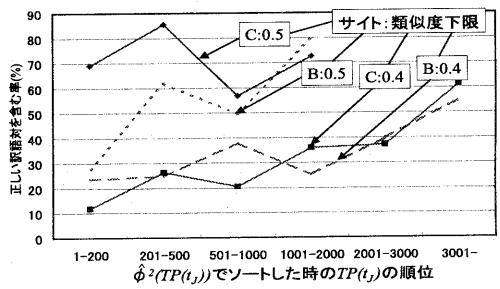


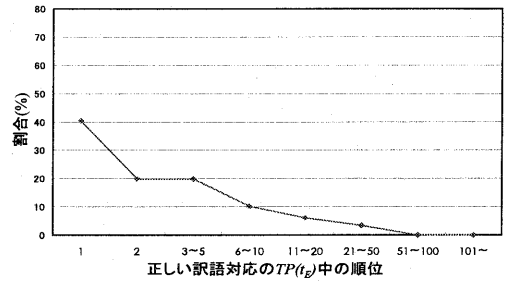
図 10: 訳語対の集合 $TP(t_J)$ に正しい訳語対を含む率の分布 (インデックス語: 日本語)

容の日本語記事の比率が小さいことが原因である．一方，サイト C において，記事間類似度の下限値 $L_d = 0.5$ の場合の「正しい訳語対を含む率」は，インデックス語が日本語の場合の方が英語の場合よりも高くなっている．この理由については，今後より詳細な分析を行う必要がある．

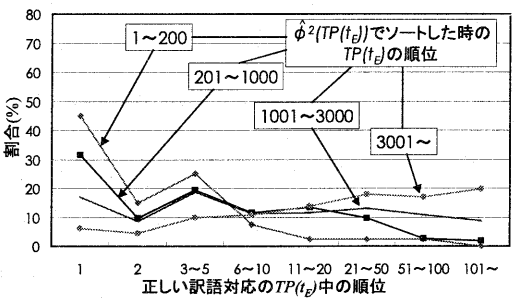
最後に，サイト A においてインデックス語を英語とした場合について，集合 $TP(t_E)$ 中の訳語対を ϕ^2 統計値の大きい順に並べ，正しい訳語対応の順位を評価した結果を図 11 に示す．まず，図 11(a) には，サイト A，記事間類似度下限 $L_d = 0.3$ の場合に，$\hat{\phi}^2(TP(t_E))$ の上位 200 位以内の集合 $TP(t_E)$ を人手で評価して抽出した訳語対 146 組について，集合 $TP(t_E)$ 中の順位の分布を示す．この結果から，上位 10 以内に約 9 割の正しい訳語対が位置していることが分かる．また，図 11(b), (c) には，記事間類似度の下限値 L_d が 0.3 および 0.5 の場合について，$\hat{\phi}^2(TP(t_E))$ の値の大きい順に集合 $TP(t_E)$ をソートした順位別に，既存の対訳辞書中の訳語対の，集合 $TP(t_E)$ 中の順位の分布を示す．この結果から，$\hat{\phi}^2(TP(t_E))$ の値が大きい方が，集合 $TP(t_E)$ 中での正しい訳語対応の順位も相対的に上位であることが分かる．これより，ϕ^2 統計値の有効性が確認できる(注8)．図 11(b) と (c) を比較すると，記事間類似度の下限値が大きければ，集合 $TP(t_E)$ 中での正しい訳語対応の順位も圧倒的に高くなることが分かる．これより，訳語対応の半自動獲得の効率が，言語横断関連報道記事検索の性能に大きく左右されることが推測される．

(注8)：本論文では，訳語候補の共起頻度のみを用いた最も簡易な統計量を用いて訳語対応の推定を行ったが，対訳コーパスを用いて連続・非連続の連語間の訳語対応を推定するためのより高性能な評価尺度 (例えば文献 [2]) を本論文の枠組に組み込むことにより，訳語対応の推定精度を改善できると考えられる．

(a) 記事間類似度下限 $L_d = 0.3$ での $\hat{\phi}^2(TP(t_E))$ 上位 200 個の $TP(t_E)$ から人手で選択した正解訳語対 146 組



(b) 既存の対訳辞書の訳語対，記事間類似度下限 $L_d = 0.3$



(c) 既存の対訳辞書の訳語対，記事間類似度下限 $L_d = 0.5$

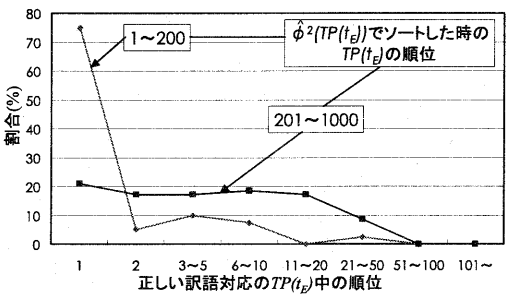


図 11: 既存の対訳辞書中の訳語対の $TP(t_E)$ における順位 (ϕ^2 統計値の降順) の分布 (サイト A)

5. 関連研究

本論文では，言語横断検索技術を援用して関連記事対を絞り込んだ結果を疑似的対訳コーパスとみなして，対訳コーパスからの訳語対応獲得手法を適用することにより訳語対応を推定した．一方，コンパラブルコーパスなどの非対訳コーパスからの訳語対応獲得においては，訳語候補の周囲の共起語の類似性を利用して訳語対応を推定することが多く (例えば，文献 [1])，そのような観点からの訳語対応の絞り込みを併用することにより，訳語対応の推定精度が改善できると期待される．ただし，従来の非対訳コーパスからの訳語対応獲得手法においては，訳語候補の周囲の共起語の頻度統計をコーパス全体から算出していた．一方，本論文の訳語対応推定の枠組では，言語横断検索技術を援用して関連記事対を絞り込んだ結果から訳語候補の周囲の共起語の頻度統計を算出することが可能であり，この絞り込みの効果を評価することが今後の重要な課題の一つである．

また，本論文の枠組では，言語横断関連記事検索技術の性能が訳語対応の半自動獲得の効率を大きく左右することになるが，今回の評価実験で採用した言語横断関連検索の手法は，最も簡易な手法の一つであると言える．今後は，二言語間の数値対応・自動生成した対訳辞書の訳語対応・発音を利用した固有名詞訳語対応 [8]，あるいは，会社名などの固有表現の訳語対応 [4] など，言語横断関連記事検索の先行研究において有効性が確認されている様々な情報を統合することにより，言語横断関連記事検索の性能を改善できると期待される．

その他の関連研究に，部分的に対訳になった文書を WWW から収集して，専門用語などの訳語対応を抽出するというものがある [7]．この種のアプローチの利点としては，一般の新聞記事などには出現することが稀であるような，特定の分野に特化した専門用語などの訳語対応の獲得に威力を発揮するという点が挙げられる．一方，WWW 上の任意の文書を収集対象とした場合，翻訳の質が安定しない可能性があり，場合によっては誤った訳語対応を抽出するという危険性もある．これに対して，本論文の枠組の大きな利点の一つとして，WWW 上の報道サイトなどを訳語対応獲得の情報源とすることから，情報源となる二言語文書の翻訳の質が十分に高く，しかも，最新の話題の二言語文書がほぼ毎日更新されるという点が挙げられる．

6. おわりに

本論文では，WWW 上の報道記事サイト等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・言い回しなどの訳語対応を半自動的に獲得する枠組を提案した．さらに，言語横断関連報道記事検索により自動収集された日英関連記事対から，半自動的に訳語対応を獲得する手法について述べ，言語横断関連報道記事検索の性能と訳語対応獲得の性能の相関について分析した．

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] M. Haruno and S. Ikehara. Two-step extraction of bilingual collocations by using word-level sorting. 電子情報通信学会論文誌, Vol. E81-D, No. 10, pp. 1103–1110, 1998.

[3] 堀内貴司, 千葉靖伸, 浜本武, 宇津呂武仁. 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, 2002.

[4] 松本賢司, 柏岡秀紀, 田中英輝. 分野固有の情報を利用した日英対訳記事コーパスの構築. 情報処理学会第 63 回全国大会講演論文集, 第 2 巻, pp. 251–252, 2001.

[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] 松本裕治, ほか. 日本語形態素解析システム『茶筌』version 2.2.8 使用説明書, 2001.

[7] M. Nagata, T. Saito, and K. Suzuki. Using the Web as a bilingual dictionary. In *Proc. Workshop on Data-driven Methods in Machine Translation*, pp. 95–102, 2001.

[8] 高橋大和, 松尾義博, 古瀬蔵. 新聞記事における日英対訳コーパスの自動構築. 言語処理学会第 5 回年次大会論文集, pp. 181–184. 言語処理学会, 1999.

盛光印刷所