翻訳知識獲得のための言語横断関連報道記事検索*

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
{takashi,chiba,hamamo,utsuro}@cl.ics.tut.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．

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．

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，言語横断関連報道記事の検索・収集・閲覧を行うシステムについて述べる．さらに，構築したシステムを用いて，日英関連報道記事対の収集を行い，収集した記事対を評価用記事集合として，言語横断関連報道記事検索の性能の評価を行った結果について述べる．

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

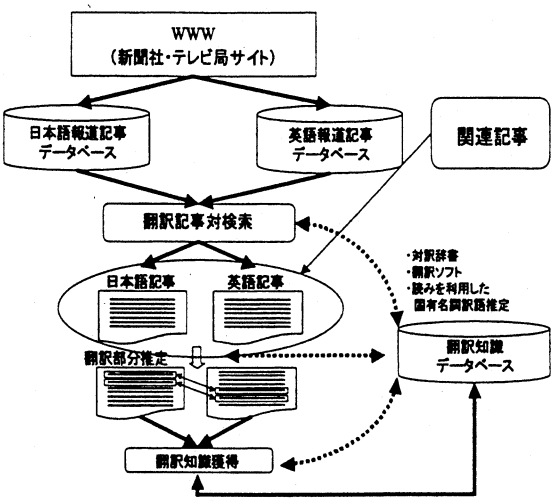

図 1: WWW からの翻訳知識獲得の流れ

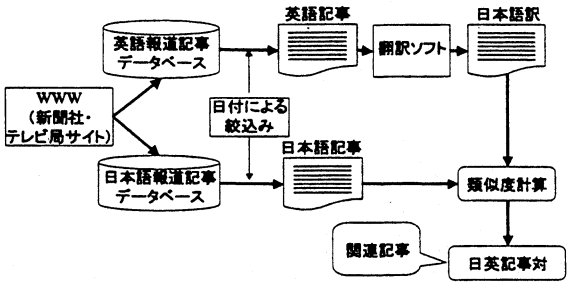

図 2: 日英関連記事対検索の流れ

得する．そして，市販の翻訳ソフト[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」 [松本，01] 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[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

*Cross-Language Retrieval of Relevant News Articles for Translation Knowledge Acquisition

[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．

[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．

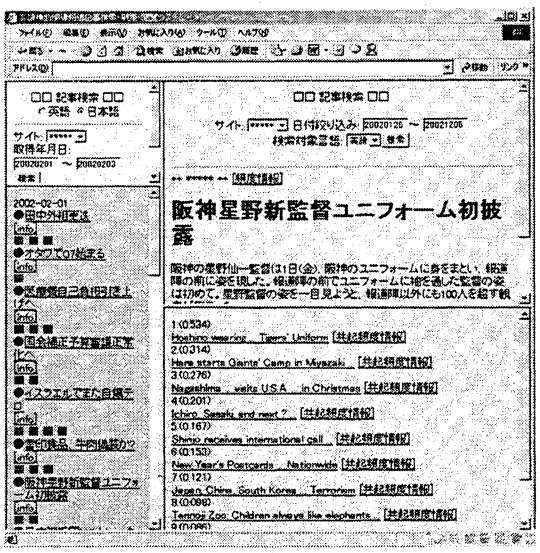

図 3: 言語横断関連報道記事検索・収集・閲覧システム

検索対象の記事を絞りこむ[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．

3 言語横断関連報道記事検索・収集・閲覧システム

本研究で作成した言語横断関連報道記事検索・収集・閲覧システムの動作画面を図 3 に示す．システムの動作の概要を以下に説明する．まず左上のフレームで，検索質問として用いる報道記事 (日本語または英語) の日付の範囲を指定する．その結果，指定された日付の範囲の記事のタイトルリストが左下のフレームに表示される．表示されたタイトルをクリックすると，右上のフレームに記事の全文が表示される (英語記事の場合は，翻訳ソフトによる翻訳結果も表示される)．記事の全文が表示されたフレーム内から，日付の範囲を指定して相手言語の関連記事を検索する．検索の結果，類似度の高い順にタイトルとその類似度が右下のフレームに表示される (ここで，“共起頻度情報” をクリックすれば，二つの記事の間の単語共起頻度が表示される)．そのタイトルをクリックすると，検索結果記事の全文が表示される (英語記事の場合は，翻訳ソフトによる翻訳結果も表示される)．表示された検索結果記事が，検索質問記事と同一の内容であるか，または密接に関連している場合には，“同一内容記事対として保存”，あるいは，“関連話題記事対として保存” のボタンをクリックすることで，関連報道記事対を保存する．左下のタイトルリストフレームは，このようにして保存された関連報道記事対を閲覧する機能を備えており，“■” をクリックすることで，言語を横断して関連報道記事を閲覧することができる．

[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．

表 1: 平均記事数・平均記事長・評価用記事対数

サイト	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	同一内容	関連話題
A	1.0	43.7	1087.3	759.9	24	33
B	14.0	89.9	3135.5	836.4	28	82
C	20.9	103.9	3228.9	837.7	28	31

4 評価

4.1 日英関連報道記事対の収集

前節で述べたシステムを用いて，日英関連報道記事対の収集を行った．まず，A～C の三種類のサイトにおける 15 日間の報道記事について，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．平均記事数においては，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，収集した日英関連報道記事対の数の内訳を表 1 の「評価用記事対数」の欄に示す．これらの記事対の収集においては，適当な英語記事を一つ選択して検索質問記事とし[4]，英語記事の前後 15 日間の範囲で関連日本語記事の検索を行った．その際は，前後 15 日間の範囲内で関連日本語記事をできるだけ網羅的に収集するために，英語記事の前後 15 日間の範囲で日付の幅を絞った検索を何回か行い，上位 40 位以内の日本語記事の範囲内で関連日本語記事を収集した．また，このようにして収集した「評価用記事対」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

[4] 詳細は後述するが，英語記事の方が記事数が少ないため，関連する日本語記事が存在する確率が高い．

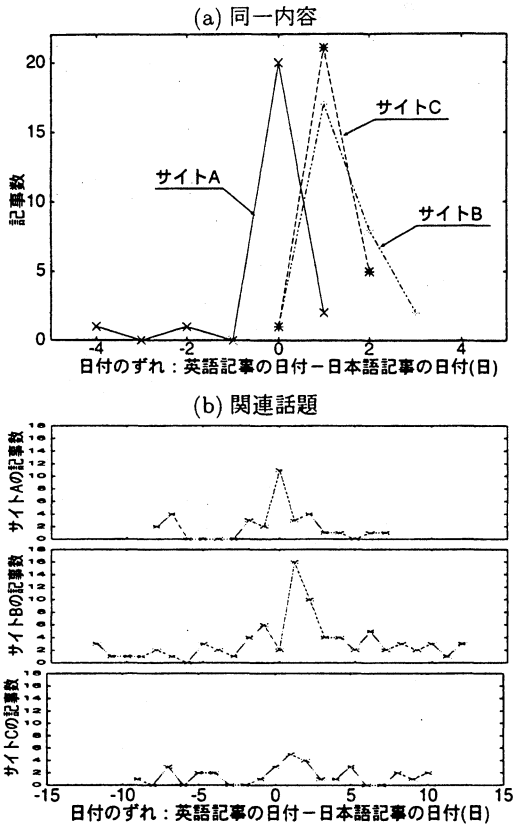

図 4: 日英関連記事間の日付のずれの分布

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．また，この集計は，前節で述べたシステムを用いて行ったが，その効率は，

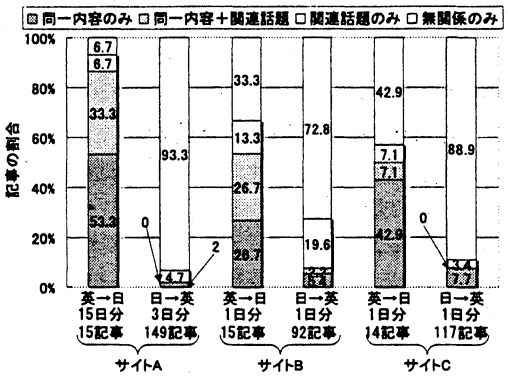

図 5: 相手言語における同一内容・関連話題・無関係記事の有無の割合

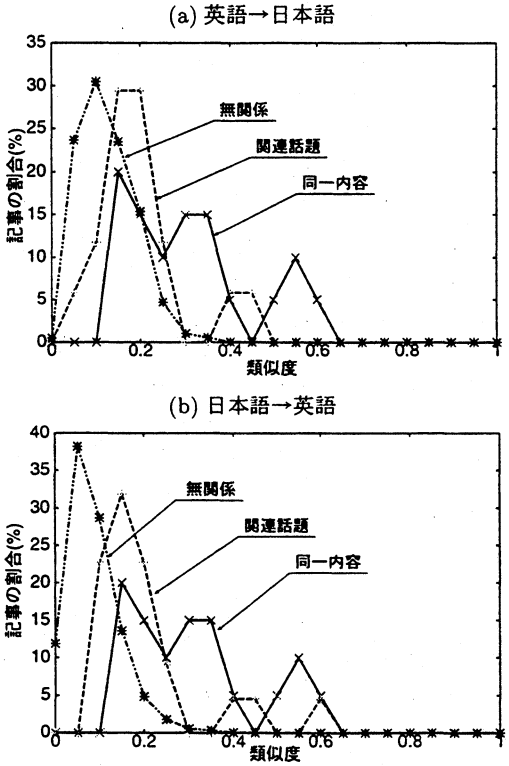

図 6: 記事間類似度の分布 (サイト C)

一検索質問記事あたり約 80 秒 (45 検索質問記事/時間) であり，作業効率としては十分高いと考えられる．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[5]．三つのグループの間で類

[5] 以降は，紙面の都合上，サイト C に対する結果のみを示すが，サイト A および B に関しても，サイト C とほぼ同等もしくはそれ以上の性能 (記事間類似度の分布および検索性能) を達成している．

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．また，英日検索と日英検索の間には極端に大きな違いは認められない．

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．評価用記事対の片方の記事を検索質問として，もう片方の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．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 ±2 日,「関連話題」: 日付のずれ ±10 日)，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,「関連話題」: 日付のずれ ±5 日) 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

$$\text{再現率} = \frac{\text{関連記事が上位 } n \text{ 位内に含まれる記事対数}}{\text{日付の範囲内に関連記事が存在する記事対数}}$$

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．また，英日検索と日英検索の比較においても，誤検索となる「無関係」記事数が少ない日英検索の方が検索性能はよい．実際に，3 節で述べたシステムを用いて日英関連報道記事対の収集を行う場合には，英語記事を検索質問として日本語関連記事を検索することが多いと考えられるが，図 7 の結果では，英日検索・日付幅最大での「同一内容」記事検索の性能は，上位 20 位以内で再現率 100%である．一方，図 5 の結果では，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する．したがって，これらの結果から，英日検索結果の上位 20 位以内から収集を行うだけで，半数以上の英語記事に対して「同一内容」の日英記事対を収集できることがわかる．

5 おわりに

本論文では，翻訳知識獲得のための言語横断関連報道記事検索の手法とその評価結果について述べた．二言語関連記事自動収集に関する関連研究としては，新聞記事を対象として，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などを用いて日英記事の対応付けを行うもの [高橋 99]，初期パラレルコーパスおよび言語横断情報検索モデル学習法を用いたブートストラップによるもの [Masuichi00]，日中関連文書収集を対象として，WWW 上の報道サイトからの候補文書収集・翻訳ソフトによる候補文書翻訳・RIDF を用いた順位付けを行うもの [Hasan01] などがある．一方，本研究では，(日本国内の)WWW 上の報道サイトから，日英関連報道記事を半自動的に効率よく収集することを目的として，言語横断関連報道記事検索・収集・閲覧システムを構築した．また，WWW 上の報道サイトにおいて内容が密接に関連した日英記事対が存在する割合，および，それらの記事対の日付の対応範囲等の調査を行った．さらに，構築したシステムを用いることにより，一定数の日英関連報道記事対を半自動的に効率よく収集できることを示した．

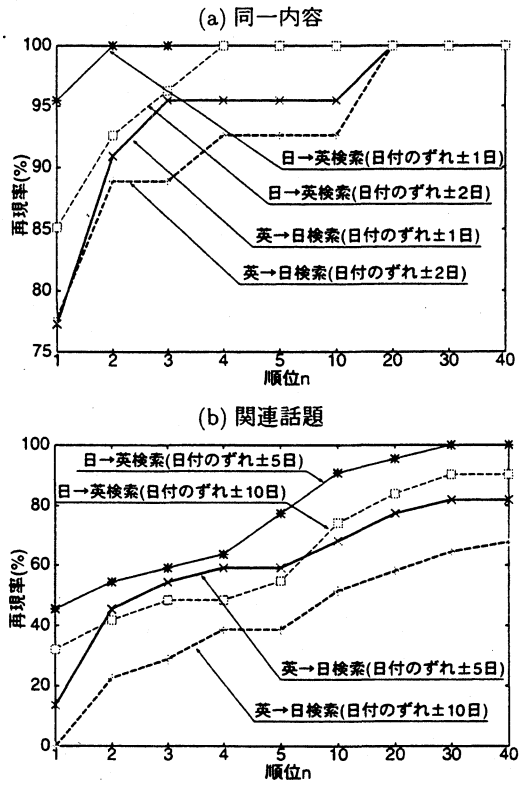

図 7: 日英関連記事検索の再現率 (上位 n 位内，サイト C)

参考文献

[Hasan01] Hasan, M. M. and Matsumoto, Y.: Multilingual Document Alignment — A Study with Chinese and Japanese, *Proc. 6th NLPRS*, pp. 617–623 (2001).

[Masuichi00] Masuichi, H., et al.: A Bootstrapping Method for Extracting Bilingual Text Pairs, *Proc. 18th COLING*, pp. 1066–1070 (2000).

[松本, 01] 松本, 他：日本語形態素解析システム『茶筌』version 2.2.8 使用説明書 (2001).

[高橋 99] 高橋, 松尾, 古瀬：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).