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*

日野 浩平　堀内 貴司　浜本 武　中山 健明
豊橋技術科学大学 工学部 情報工学系
{hino,takashi,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．

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 [堀内03]．

この一連の枠組において，特に本論文では，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ためのユーザインタフェースについて述べる．このユーザインタフェースは，多数の訳語対応の候補を構造化しておき，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機能を持つ．さらに，このユーザインタフェースを用いて，WWW上の日英関連報道記事から実際に訳語対応を獲得した結果についても述べる．

*User Interface for Translation Knowledge Acquisition from Japanese-English Relevant News Articles

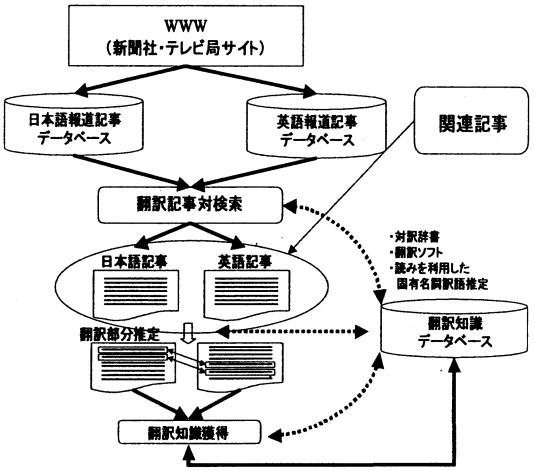


図 1: 日英関連報道記事からの翻訳知識獲得の流れ

2 日英関連報道記事からの訳語対応の獲得

2.1 訳語対応の推定

本節では，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 [堀内03] の概要について述べる．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．

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．

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

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．

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

本論文では，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，この集合 RC_{EJ} から訳語対応を推定し，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ を求める [堀内 03]．ここで，訳語対応推定の対象となる英語連語または単語の品詞列としては任意のものを，また，日本語連語または単語の品詞列としては，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する[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．

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

実際の評価実験では，$L_f^E = L_f^J = 3, U_l^E = U_l^J = 5$ という条件を設定している．また，[堀内 03] では，この集合 RC_{EJ} から訳語対応を推定する方法として，この集合 RC_{EJ} 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，この集合 RC_{EJ} 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評価を行っている．特に，この集合 RC_{EJ} を疑似的な対訳コーパスとみなす場合には，関連する記事組 $\langle d_E^i, D_J^i \rangle$ 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} を設ける．実際の評価実験では，$L_f^{EJ} = 2$ という条件を設定している．

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

2.2 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日英関連報道記事対の集合 RC_{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が共通の訳語組をまとめる．
2. 英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．

[1] 「茶筌」の品詞体系では，接尾辞は名詞に含まれる．

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，t' が t の一部を構成するという関係を $t \succeq t'$ で記述する[2]．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 $(t'_E \not\succeq t_E)$ 場合に，以下の手順で訳語組の集合 $TP_c(t_E)$ 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．

$$TP_c(t_E) = \left\{ \langle t''_E, t_J \rangle \mid \forall t'_E (\neq t_E)\ t'_E \not\succeq t_E,\ t_E \succeq t''_E \right\}$$

集合 $TP_c(t_E)$ 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 (ただし，頻度下限および構成単語数の上限を満たす) 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[3]．

次に，各集合 $TP_c(t_E)$ 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 $corr_{EJ}(TP_c(t_E))$ とする．

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

そして，全ての集合 $TP_c(t_E^1)$, ... , $TP_c(t_E^m)$ を $corr_{EJ}(TP_c(t_E))$ の値の降順に並べ，先頭から順に各集合 $TP_c(t_E)$ の要素を人手で調べていくこととする．また，各集合 $TP_c(t_E)$ の要素である訳語対を人手で調べる際には，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ の降順に調べることとする．

3 訳語対応獲得のためのユーザインタフェース

3.1 ユーザインタフェースの機能

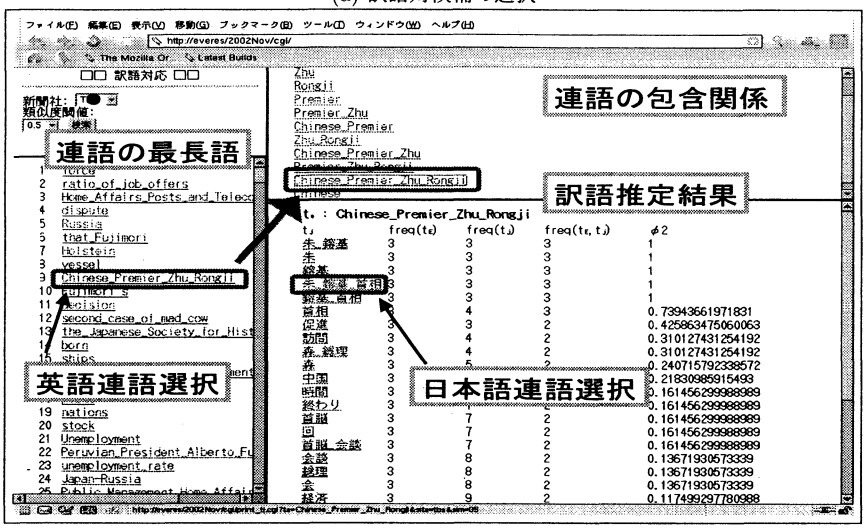
次に，実際の動作例 (図 2) を用いて，本研究で作成した，訳語対応獲得のためのユーザインタフェースの機能について説明する．このインタフェースでは，前節の手順により構造化された訳語対応推定結果を走査しながら，必要に応じて，言語横断関連報道記事検索により自動収集された日英関連記事対を閲覧することにより，正しい訳語対応を効率よく選び出す．

まず，図 2(a) 左側のフレームには，全ての集合 $TP_c(t_E^1), \ldots, TP_c(t_E^m)$ を訳語対応推定値の最大値 $corr_{EJ}(TP_c(t_E))$ の降順に整列し，各集合のインデックス語 t_E を表示している．ここで，インデックス語 t_E として，“Chinese Premier Zhu Rongji” を選んだとすると，この t_E もしくはその部分単語列となる英語連語もしくは単語が右上のフレームに表示される．

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

[3] 現時点では，インデックス語を英語として，訳語対応獲得のユーザインタフェースを実装しているが，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獲得のユーザインタフェースを実装し，訳語対応推定結果の評価を行うことができる．

(a) 訳語対候補の選択



(b) 訳語対候補を含む英語記事・日本語記事の閲覧

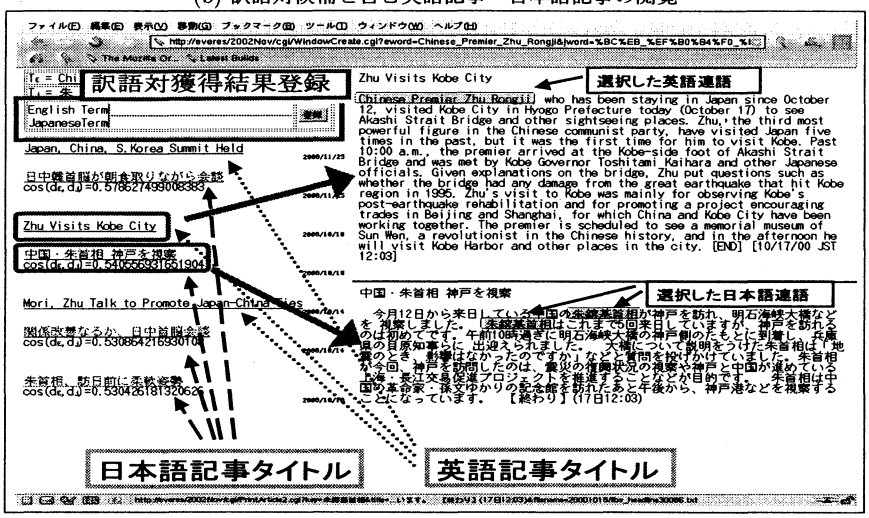


図 2: 訳語対応獲得のためのユーザインタフェースの動作例

次に，これらの英語連語もしくは単語の中から，訳語対応獲得の対象とすべき適切な語 (この動作例では，“Chinese Premier ZhuRongji”) を選択すると，右下のフレームに，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ の降順に頻度 $freq(t_E)$，$freq(t_J)$，$freq(t_E, t_J)$，および，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ を示す (図 2 の動作例では，訳語対応推定値 $corr_{EJ}(t_E, t_J)$ として ϕ^2 統計を用いている)．この例の場合には，$corr_{EJ}(t_E, t_J)$ の値が最も大きい (ものの一つ)「朱 鎔基 首相」がインデックス語 t_E の正しい訳語に最も近い語となっている．

次に，作業者は，図 2(b) に示すように，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 $\cos(d_E, d_J) \geq L_d$ (図 2(a) 左上のフレームにおいて最初に指定しておく) 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．図 2(b) の日英記事組閲覧のウィンドーを起動する際には，図 2(a) の訳語組候補表示・選択のウィンドーにおいて，右下のフレームの日本語連語もしくは単語のリストの中から，適当に一つを選択してクリックする．図 2 の場合には，図 2(a) 右下のフレームで「朱 鎔基 首相」を選択することにより，英語連語 t_E として “Chinese Premier Zhu Rongji”，日本語連語 t_J として「朱 鎔基 首相」がそれぞれ選択されて，これらを含む英語記事および日本語記事のタイトルが，図 2(b) のウィンドーの左半分に表示される．具体的には，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

表 1: 記事の日数・記事数・平均記事長

総日数		総記事数		記事間類似度 0.4 以上の記事数		一日の 平均記事数		一記事の平均 記事長 (byte)	
英語	日本語	英語	日本語	英語	日本語	英語	日本語	英語	日本語
562	578	607	21349	190	377	1.1	36.9	1087.3	759.9

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．これらの記事のタイトルをクリックすると，英語記事および日本語記事の本文がそれぞれ右上と右下のウィンドーに表示され，特に，選択された英語連語 t_E および日本語連語 t_J の部分は，色付で強調されて表示される．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．

図 2(b) の動作例の場合は，“Chinese Premier Zhu Rongji” および「朱 鎔基 首相」を含む英語および日本語記事の本文を閲覧することにより，これらの連語を含む記事がほぼ同一の内容であることが分かる．さらに，この場合，“Chinese Premier Zhu Rongji” の完全な翻訳は，「中国の朱鎔基首相」となっているが，日本語側での連語を認定する基準として，品詞を名詞または動詞に制限していることから，“Chinese Premier Zhu Rongji” の完全な訳語対応を獲得することはできない．この場合には，訳語対応としては，“Premier Zhu Rongji” および「朱 鎔基 首相」の組を獲得することができる．

3.2 使用例

表 1 に示す日数および記事数の記事集合に対して，記事間類似度が 0.4 以上の記事組を収集し，2.1 節で述べた方法により訳語対応を推定した [堀内 03]．そして，2.2 節で述べた，訳語対応推定値の最大値 $corr_{EJ}(TP_c(t_E))$ の上位 200 個の $TP_c(t_E)$ に対して，本論文の一人である大学院生が本論文のユーザインタフェースを用いて，訳語対の半自動獲得を行った．その結果，2.1 節で述べた方法のうち，疑似的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 (ϕ^2 統計) を適用して訳語対応を推定した場合は，約 5.4 時間で 164 組の訳語対 (内，53 組=32.3%が既存の対訳辞書 (英辞郎 Ver.37: 85 万語) に含まれず) が獲得でき，

表 2: 訳語対獲得の例

対訳辞書に訳語対として存在	
Hansen's disease	ハンセン病
mad cow disease	狂牛病
Cabinet Office	内閣府
Yasukuni Shrine	靖国神社
対訳辞書に訳語対として存在せず	
conference on Afghan reconstruction	アフガン復興会議
New job offers	新規求人
Prime Minister Yoshiro Mori	森総理大臣
Kato faction	加藤派

コンパラブルコーパスからの訳語対応推定手法を適用して訳語対応を推定した場合は，約 3.7 時間で 116 組の訳語対 (内，63 組=54.3%が既存の対訳辞書に含まれず) が獲得できた [堀内 03]．いずれの場合も，一時間あたり 30 組程度 (内，10~15 組程度が既存の対訳辞書に含まれず) の訳語対が獲得できており，作業者に対して高い専門知識が必要とされないことを考慮すると，高い作業効率であると言える．獲得された訳語対のうち，既存の対訳辞書に訳語対として存在する組と存在しない組の例を表 2 に示す．

4 おわりに

本論文では，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ためのユーザインタフェースの枠組を提案し，その有用性について述べた．

参考文献

[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)．

[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，Vol. 2002, No. (2002-NL-150), pp. 191–198 (2002)．

[堀内 03] 堀内貴司，日野浩平，浜本武，中山健明，宇津呂武仁：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 (2003)．

[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).