2004－NL－162　(10)
2004／7／15

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

木田　充洋† 　宇津呂武仁† 　日野　浩平‡ 　佐藤　理史†

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

我々はこれまで，二言語コーパスからの翻訳知識獲得のアプローチの一つとして，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，報道内容が密接に関連した日本語記事および英語記事を収集し，そこから翻訳知識を獲得するアプローチを提案し，その有効性を示してきた．しかし，それらの報道記事においても，言語資源の規模の大きさには限界があり，出現頻度の少ないタームについては，訳語候補を列挙するにとどまり，訳語候補の有効な順位付けが難しいという点が問題である．そこで，本稿では，出現頻度の少ないタームに対する訳語候補の順位付けを効果的に行うために，ウェブ検索エンジンを用いて各タームの出現する日英非対訳文書を収集し，訳語候補順位付けの情報源とする手法を提案する．さらに，評価実験において，報道記事中に低頻度で出現するタームについて，報道記事を用いる方法よりも高精度の順位付けを行うことができることを示す．

Estimating Bilingual Term Correspondences from Japanese and English Documents

Re-estimation using Monolingual Web Documents

Mitsuhiro KIDA†, Takehito UTSURO†, Kohei HINO‡, and Satoshi SATO†

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

In the framework of bilingual lexicon acquisition from cross-lingually relevant news articles on the Web, it is relatively harder to reliably estimate bilingual term correspondences for low frequency terms. Considering such a situation, this paper proposes to complementarily use much larger monolingual Web documents collected by search engines, as a resource for reliably re-estimating bilingual term correspondences. We experimentally show that, using a sufficient number of monolingual Web documents, it is quite possible to have reliable estimate of bilingual term correspondences for those low frequency terms.

1. はじめに

我々はこれまで，二言語コーパスからの翻訳知識獲得のアプローチの一つとして，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，報道内容が密接に関連した日本語記事および英語記事を収集し，そこから翻訳知識を獲得するアプローチを提案し，その有効性を示してきた [3], [6]．このアプローチは，情報源となるコーパスを用意する段階においては，対訳コーパスを用意するために必要となるような大きなコストを必要としない．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，従来のコンパラブルコーパスからの翻訳知識獲得のアプローチと比較して，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．

このアプローチでは，情報源となる報道記事中に一定頻度以上出現するタームについては，比較的安定して訳語対応等の翻訳知識の獲得が行える．しかし，出現頻度の少ないタームについては，訳語候補を列挙するにとどまり，訳語候補の有効な順位付けが難しいという点が問題となっていた．このような状況をふまえ，本稿では，出現頻度の少ないタームに対する訳語候補の順位付けを効果的に行うために，ウェブ検索エンジンを用いて各タームの出現する日英非対訳文書を収集し，訳語候補順位付けの情報源とするというアプローチをとる．訳語対応を推定する手法としては，タームの出現する文から構成した文脈ベクトルの類似性を用いる方法，および，タームの出現する文書の類似性を用いる方法を評価した．用いる非対訳文書としては，検索エンジンにより得られるサイト内の文書を用いる方法，および，検索エンジンにより検索した結果のサマ

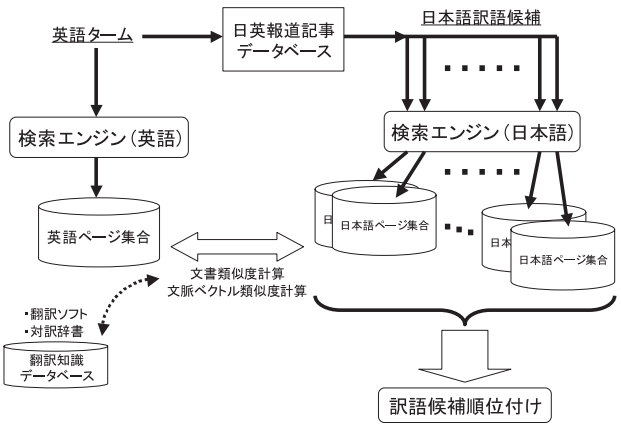

図 1　ウェブ検索エンジンを利用した訳語推定の流れ

リページを用いる方法を比較した．言語を横断して文脈ベクトルあるいは二言語文書の類似性を測定するための情報源の性能としては，対訳辞書および翻訳ソフトを比較した．ウェブ検索エンジンにより収集される日英非対訳文書においては，二言語間における内容の関連性が低いため，関連報道記事と比較しても，訳語対応推定は容易ではないと予想される．しかし，評価実験の結果では，報道記事を用いた訳語対応推定よりも精度は低下するものの，ある程度の精度は保っていた．また，報道記事において候補の順位付けが難しいとされる，報道記事中に低頻度で出現するタームについて，報道記事を用いる方法よりも高精度の順位付けを行うことができた．これらのことから，ウェブ検索エンジンにより収集される日英非対訳文書の有効性が確認できた．

2. ウェブ検索エンジンにより収集された非対訳文書を用いた訳語推定

2.1 概　　要

訳語対応推定のタスクは，大きく次の二つのサブタスクに分けて考えることができる．i) 対象の英語タームに対して日本語訳語候補を収集する，ii) 訳語候補の順位付けを行う．本稿では，i) の過程については，関連報道記事から得られる訳語候補 [3] を用いることとし，ウェブ検索エンジンにより収集される日英非対訳文書を利用して ii) の訳語候補順位付けを行う．この流れを図 1 に示す．まず，ウェブ検索エンジンを用いて英語タームおよび日本語訳語候補をを含む文書をそれぞれ収集する．次に，収集した文書から文書ベクトルを作成する．ここで，日本語文書はそのままベクトル化し，英語文書は，翻訳ソフトあるいは対訳辞書を用いて日本語訳に変換した後，その日本語訳をベクトル化する．最後に， 英日文書間で文書ベクトルを用いて訳語対応推定を行う．訳語対応を推定する手法としては，タームの出現する文から構成した文脈ベクトルの類似性を用いる方法，および，タームの出現する文書の類似性を用いる方法を評価した．

2.2 ウェブヒット数による訳語候補絞り込み

実際に訳語対応推定を行う前に，ウェブ検索エンジンより得られるヒット数を用いた訳語候補の絞り込みを行う．ここでは，英語ターム t_E と日本語ターム t_J が訳語の関係にある場合には，それらのタームのヒット数 $h(t_E)$ と $h(t_J)$ の間に一定の相関があると考え，$h(t_E)$ の範囲によって，経験的に，$h(t_J)$ の下限 h_L および上限 h_U を定める．

$$h_L < h(t_J) \leqq h_U$$

今回の実験では，下限 h_L および上限 h_U を以下のように定めた(注1)．

（1） $0 < h(t_E) \leqq 100$ の場合，
$h_L = 0,\ h_U = 10,000 \times h(t_E)$

（2） $100 < h(t_E) \leqq 20,000$ の場合，
$h_L = 0.05 \times h(t_E),\ h_U = 1,000,000$

（3） $20,000 < h(t_E)$ の場合，
$h_L = 1,000,\ h_U = 50 \times h(t_E)$

2.3 非対訳文書の収集・ベクトル化

英語ターム t_E および日本語ターム t_J をそれぞれクエリとして，検索エンジンにより文書を収集する．得られた文書集合をそれぞれ $D(t_E)$，$D(t_J)$ とする．本論文では，検索エンジンにより得られる以下の 2 通りの文書のいずれかを用いている．

- 検索エンジンにより得られた各ページの本文全体.
- 検索エンジンにより検索した結果のサマリページにおいて，各サイトの URL ごとに得られる要約テキストの文書.

次に，得られた文書から html タグを除去し，英語文書は翻訳ソフト (オムロン社製「翻訳魂」) あるいは対訳辞書 (英辞郎 Ver.37，85 万語) により日本語訳に変換する．対訳辞書を用いる場合は，英語単語もしくは 5 単語長以下の英語連語に対して得られる全訳語候補を列挙し，これを日本語訳とする．これらの日本語文書に対して，日本語形態素解析システム「茶筌」(注2)により形態素列への分割を行う．そして，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て文書の頻度ベクトルを作成する．

2.4 訳語候補の順位付け

2.4.1 文脈ベクトルの類似性を利用する方法

文脈ベクトルの類似性を用いて訳語対応推定を行う場合は，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、英語文書集合 $D(t_E)$ 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加算して，t_E に対する文単位の文脈頻度ベ

(注1)：ウェブ検索エンジンとしては，英語は AltaVista (http://www.altavista.com/，日本語は goo (http://www.goo.ne.jp/) を用いた．検索エンジンを用いた文書収集は 2004 年 1 月に行ったが，この下限・上限の決め方は，利用する検索エンジンおよび文書収集を行う時期によって変化する可能性があると思われる．なお，今回の実験では，この下限・上限により，一英語タームあたり 50 個の訳語候補を平均 43.5 個に絞り込んだ.

(注2)：http://chasen.aist-nara.ac.jp/

クトル $cv_{trJ}(t_E)$ を構成する．同様に，日本語文書集合 $D(t_J)$ 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 $\cos(cv_{trJ}(t_E), cv(t_J))$ を訳語対応推定値 $corr_{EJ}(t_E, t_J)$ とする．

2.4.2 文書間の類似性を利用する方法

文書の類似性を用いて訳語対応推定を行う場合は，まず，文書類似度計算を安定して行うために，文書の一部を削除してテキストサイズの正規化を行い，ウェブ上の報道記事 [3] と同等のサイズ(注3)となるようにする．この際，文書中で，それぞれ，t_E あるいは t_J を含む部分は削除しないものとする．次に，t_E を含む英語文書の集合を $D(t_E)$，t_J を含む日本語文書の集合を $D(t_J)$ として，$D(t_E)$ 中の文書 d_E，および，$D(t_J)$ 中の文書 d_J との間で，文書間類似度を計算する．文書間類似度としては，d_E の日本語訳文書の頻度ベクトル $v_{trJ}(d_E)$ と d_J の頻度ベクトル $v(d_J)$ の間の余弦 $\cos(v_{trJ}(d_E), v(d_J))$ を用いる．そして，この文書間類似度が下限値 L_d 以上となる文書組の集合を $DD(t_E, t_J, L_d)$ とする．

$$DD(t_E, t_J, L_d) = \left\{ < d_E, d_J > | \, d_E \in D(t_E), d_J \in D(t_J), \cos(v_{trJ}(d_E), v(d_J)) \geqq L_d \right\}$$

最後に，全文書組数に対するこの文書組数の割合

$$\frac{|DD(t_E, t_J, L_d)|}{|D(t_E)||D(t_J)|}$$

を求め，これを訳語対応推定値 $corr_{EJ}(t_E, t_J)$ とする．

3. 実験および評価

3.1 ウェブから収集した文書

今回の実験では，評価用の英語タームとしては，文献 [3] の「文書数と訳語対応推定性能の相関」の評価において，頻度 10 以上の英語タームから，ϕ^2統計値 で降順に上位 100 個を選定したもの，および，別途選定した 29 個の英語タームセットを用いる(注4)．本稿では，これらの各英語タームにつき，文献 [3] の ϕ^2 統計値の上位 50 個の日本語訳語候補 (2.2 節のヒット数による絞り込み後は平均 43.5 個) を用いて訳語候補順位付け手法の評価を行う(注5)．

2.3 節および 2.4 節で述べたように，訳語対応推定においては，英語文書翻訳法として翻訳ソフト・対訳辞書の二通りがあり，訳語対応推定尺度として，文脈ベクトルを用いる場合と文書間類似度を用いる場合の二通りがある．したがって，訳語対応推定の方法は合計四通りとなる．3.3 節で述べるように，この四通りのうち最も性能がよい対訳辞書・文脈ベクトルの組を用いる方法において，情報源として用いる文書数と訳語対応推定精度の相関を調べた結果では，文書数を 50~1,000 の範囲で変化させても，訳語対応推定精度が大幅に下がることはなかった．そこで，より大きい評価タームセットである 100 英語タームを用いた評価実験においては，一タームあたり 100 文書を収集して，対訳辞書・文脈ベクトルの組を用いた訳語対応順位付け手法の評価を行った．この場合，一タームあたりの平均文書数は，英語で 74.4，日本語で 83.9 となった．

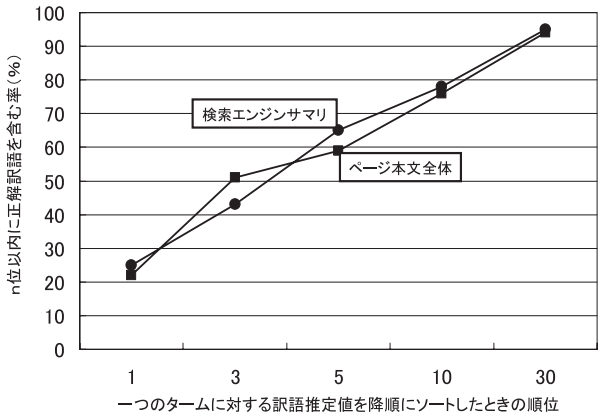

図 2　文脈ベクトルを用いた訳語対応推定: ページ本文全文と検索エンジンサマリの比較 (100 ターム)

3.2 ページ本文全文と検索エンジンサマリの比較

検索エンジンにより収集できるページの本文を用いる訳語候補の順位付けでは，文書収集に膨大な時間がかかる．また，非常に大量の文書を用いるため，保存に膨大な容量を必要とする．そこで，収集できるページの本文を用いるかわりに，2.3 節で述べたように，検索エンジンにより検索した結果のサマリページにおいて，各サイトの URL ごとに得られる要約テキストの文書を用いる方法を評価し，両者を比較した．ここで，サマリページを用いる方法の利点としては，タームを含むそれぞれのページにアクセスしないため，文書収集にかかる時間が大幅に短縮できる点が挙げられる．

この実験では，100 タームを評価用英語タームとして，対訳辞書・文脈ベクトルの組を用いた訳語対応推定において，情報源として収集できるページの本文を用いる方法，および，検索エンジンにより検索した結果のサマリページを用いる方法を比較した．情報源として用いる文書数は，本文，サマリともに各タームにつき最大 100 文書とした．ただしヒット数が 100 未満のタームが含まれていたため，実際には一タームあたりの平均文書数は，英語で 76.0，日本語で 86.4 となった．以上の 2 種類の文書において，訳語候補の順位付けを行った結果を図 2 に示す．横軸は正解訳語が含まれる順位，縦軸は上位 n 位以内に正解訳語 (この実験では，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に対応している．

(注3)：英語文書 200~600 単語，日本語文書 1,500~4,000 バイト．

(注4)：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．

(注5)：文献 [3] においては，英語タームの頻度分布，および，ϕ^2 統計値の値によって，全英語タームの集合を 9 通りに分類し，各分類ごとに 100 ターム程度の評価用英語タームを選定し，訳語対応推定の評価を行なっている．これらの 9 種類のタームセットについても，本稿の手法による訳語候補順位付け手法の評価を行なっており，以下で述べる評価結果と同等の性能が確認できている．

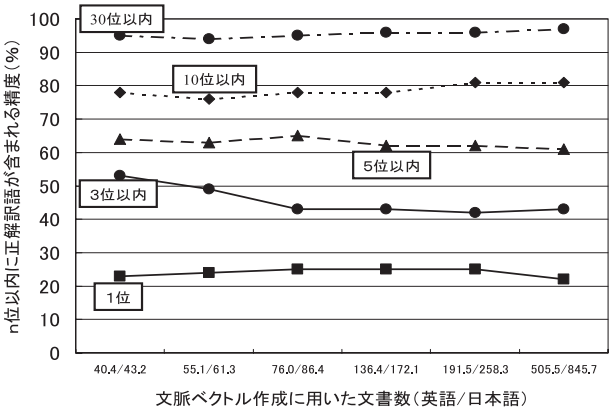

図 3　文書数と訳語対応推定精度の相関 (100 ターム)

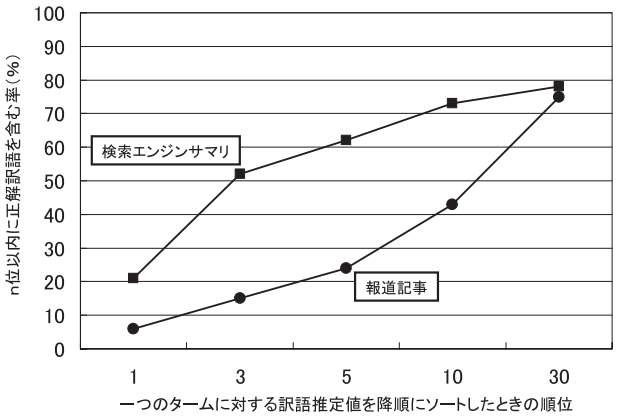

図 4　少量の報道記事を用いた訳語候補順位付けとの比較 (100 ターム)

この結果から，検索エンジンにより検索した結果のサマリページを用いる手法は，タームを含むページ本文を用いる手法とほぼ同等の性能を保っていることが分かった．また，文書の収集から訳語候補の順位付けを行うまでにかかる計算時間は，約 20 分の 1 に短縮できた．このことから，タームを含むページの本文を用いるかわりに，検索エンジンにより検索した結果のサマリページを用いることで，比較的短時間で効率的に，これまでと同様の精度で訳語対応推定を行うことができることが分かった．

3.3　文書数と訳語対応推定精度の相関

100 タームを評価用英語タームとして，対訳辞書・文脈ベクトルの組を用いた訳語対応推定において，情報源として用いる文書数と訳語対応推定精度の相関を評価した．文書数を 50〜1,000 の範囲で変化させて，上位 $n(=1,3,5,10,30)$ 位以内に正解訳語が含まれる英語タームの割合をプロットした結果を図 3 に示す．

横軸には，最大文書数を 50，70，100，200，300，1000 とした場合の実際の平均文書数を英語・日本語ごとに示す．この結果から分かるように，最大文書数を減少させても大幅な精度の低下は観測されなかった．したがって，今回の実験では，一タームにつき 50〜100 文書程度を収集すればほぼ十分であると言える．

3.4　報道記事を用いた訳語対応推定との比較

報道記事において出現頻度の低いタームは，報道記事を情報源とした訳語候補順位付けではよい性能を得られないことが分かっている [3]．このような場合には，ウェブ上の日英単言語文書を情報源として用いることにより，訳語候補順位付けの性能を効果的に改善できるのではないかと考えられる．そこでここでは，3.1 節で述べた評価用 100 英語タームに対して，少量の報道記事から得られる日本語訳語候補(注6) について，ウェブ上の日英単言語文書を情報源として訳語候補順位付けを行った．報道記事およびウェブ上の日英単言語文書を用いた訳語候補順位付け結果を図 4 に示す．図中の訳語候補順位付け精度は，上位 n 位以内に正解訳語が含まれる英語タームの割合に対応している．なお，この実験ではウェブ上の日英単言語文書として，検索エンジンにより検索した結果のサマリページを用いている．

この結果から分かるように，ウェブ上の日英単言語文書を用いた場合では，低頻度のタームについて報道記事を用いた場合と比べて高性能で訳語候補の順位付けを行うことができた．これは，報道記事では出現頻度が低く，多様な内容の文書を収集することができないのに対して，検索エンジンを用いる場合では，ウェブ上の多様な内容の文書を収集することができたためであると考えられる．このことから，報道記事において出現頻度の低いタームの訳語候補順位付けにおいて，ウェブ上の日英単言語文書を情報源として用いることが有用であることが確認できた．

3.5　文書の翻訳法および訳語候補順位付け尺度の比較

評価用 29 英語タームについて，翻訳ソフト・対訳辞書の二通りの方法による英語文書の翻訳法と，文脈ベクトル・文書間類似度の二通りの訳語対応推定尺度の計四通りについて訳語候補順位付け精度を比較した結果を図 5 に示す．図中の訳語候補順位付け精度は，上位 n 位以内に正解訳語が含まれる英語タームの割合に対応している．なお，文書間類似度を用いた訳語対応推定尺度においては，文書間類似度の下限値 L_d としていくつかの値を評価したが，0.2 と 0.3 の性能が比較的高く，その中でもやや高い性能を示した $L_d = 0.2$ の場合の結果を示す．

この結果から分かるように，翻訳ソフトと対訳辞書の比較では，対訳辞書の方が高い性能を示した．また，文脈ベクトルと文書間類似度では，文脈ベクトルを用いた訳語対応推定尺度の方が高い性能を示した．特に，翻訳ソフトと対訳辞書の比較においては，日英関連報道記事を用いた訳語候補順位付け [4] において，翻訳ソフトの方が圧倒的に高い性能を示したのに対して，本稿では逆の結果が得られた．

(注6)：報道記事を用いる手法 [3] では，まず対象の英語タームが頻度 10 回となるように英語記事を選定した．次に，各英語記事の前後 2 日以内の日本語記事を選定し，これらの記事組を情報源として訳語対応推定を行った．ここで，ϕ^2 統計値の上位 50 個の日本語訳語候補は，3.1 節で述べた日本語訳語候補とは一致していない．また，2.2 節のヒット数による絞り込み後は平均 24.8 個となっている．

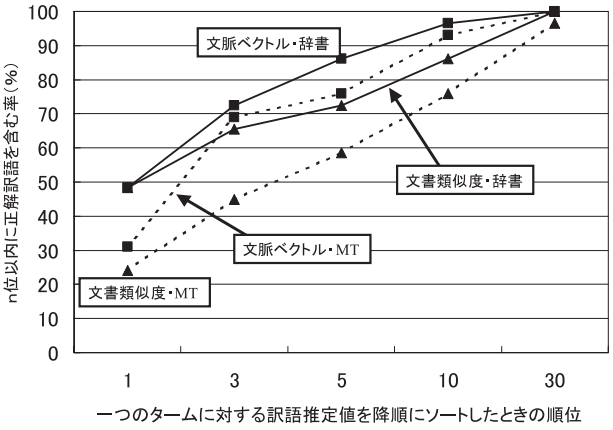

図 5　翻訳ソフトと対訳辞書の精度比較 (29 ターム, 文書類似度の閾値 $L_d = 0.2$)

ここで，翻訳ソフトによる翻訳と対訳辞書による翻訳を比較すると，翻訳ソフトによる翻訳では，複数の意味を持つ語句に対して文脈を考慮した訳語選択が行われるのに対して，対訳辞書による翻訳では，全ての訳を列挙する点が大きく異なっている．日英関連報道記事を用いた訳語候補順位付けの場合は，日付の近い二言語記事の間でお互いに関連した内容が含まれる可能性が高いために，一意に訳語選択を行う翻訳ソフトによって記事翻訳を行い，相手言語の関連記事を検索するという方法が適していると考えられる．一方，ウェブ検索エンジンにより収集した非対訳文書の場合は，二言語間で関連した内容の文書が存在する割合がかなり低い．したがって，翻訳ソフトにより一意の訳語選択を行うよりは，対訳辞書により全ての訳語を列挙した方が，分野が近い相手言語文書と他分野の相手言語文書を識別する性能が高くなると考えられる．

3.6 考　　察

本節では，ウェブ上の日英単言語文書を用いた訳語候補順位付けにおいて，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原因を分析する．さらに，報道記事を用いた場合の訳語候補順位付けとの比較を行なう．考察対象の英語タームとしては，文献 [3] で用いたものから，無作為に 300 タームを選択した．これらの 300 タームに対して，ウェブ上の日英単言語文書を用いて訳語候補順位付けを行なったところ，図 2 とほぼ同等の性能であった．

まず，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原因を大きく以下の六通りに分類し，その内訳を表 1 に示す．

（1） 正解訳語と同義語が一位
（2） 正解訳語の部分列が上位
（3） 関連語が上位
（4） 報道記事における訳語候補に正解訳語が含まれない
（5） 英語タームのヒット数 10 未満
（6） ウェブヒット数による絞り込み

表 1　訳語候補順位付けの誤り原因の分析

誤り原因	割合 (%)
正解訳語と同義語が一位	4.6
正解訳語の部分列が上位	30.0
関連語が上位	49.3
報道記事における訳語候補に正解訳語が含まれない	7.4
英語タームのヒット数 10 未満	4.6
ウェブヒット数による絞り込み	4.1

このうち,「正解訳語と同義語が一位」は，訳語対応推定値の一位が正解とみなしてもよいと考えられる.「正解訳語の部分列が上位」は，例えば,「脱税事件」が正解訳語の場合に，その部分列「脱税」が上位となるような場合である.「関連語が上位」は，例えば,「燃料電池車」が正解訳語の場合に，同一文書内で共起する関連語「水素」「次世代」「エネルギー」などが上位となる場合である.「英語タームのヒット数 10 未満」は，英語タームのヒット数が極端に小さく，十分な数の英語文書が収集できないという場合である．また,「ウェブヒット数による絞り込み」は，英語タームと日本語正解訳語の間でヒット数の大きさが極端に異なるため，2.2 節の「ウェブヒット数による訳語候補絞り込み」により，正解訳語が削除されたという場合である．

ここで，報道記事を用いた訳語候補順位付けに対して，ウェブ上の日英単言語文書を用いた訳語候補順位付けが持つ利点の一つとして，ウェブ上の多様な文書を収集することにより，個々の訳語候補について，その文脈特性に応じた多様な文脈ベクトルが構成でき，訳語候補順位付けがより高精度に行なえる可能性があるという点が挙げられる．特に，上述の誤り原因の内訳においては,「関連語が上位」が誤りの約半数を占めるが，これらのうちの多くは，例えば，訳語候補ごとにできるだけ異なる文書集合を収集した上で訳語対応推定を行うというアプローチにより，改善が可能であると考えている．実際に，これらの誤り例の中には，正解訳語が一位とはならなかったものの，報道記事を用いた訳語候補順位付けとの比較においては，正解訳語の順位が改善し，逆に関連語である誤りの訳語候補の順位が下がったという事例が観測されている．これらの例を表 2 に示す．これらの例においては，いずれも，報道記事中で正解訳語と共起した関連語の順位が下がり，正解訳語の順位が上がるという結果が得られている．

4. 関 連 研 究

従来より，コンパラブルコーパスを用いた訳語対応推定の研究 (例えば，文献 [2]) においては，タームの周囲の文脈から構成した文脈ベクトルの類似性により訳語対応を推定する手法がよく用いられる．これらの方法では，あるタームに対する訳語の候補をいかにして収集するかが問題となるが，コーパスの規模を制限したり訳語候補を単語に限定したりして，訳語候補の探索範囲を実際に計算可能

表 2　報道記事およびウェブ上の日英単言語文書における訳語候補順位の改善例

英語ターム	正解日本語訳語	その他の日本語訳語候補
Asia-Pacific Economic Cooperation forum summit	アジア太平洋経済協力会議 (11 位 →3 位)	日米韓首脳 (4 位 →13 位) 中曽根弘文参院議員 (8 位 →38 位) 出馬辞退 (6 位 →37 位)
Political Ethics Hearing Committee	政治倫理審査会 (8 位 →3 位)	不要論 (4 位 →30 位) ロシア課 (5 位 →26 位)
Civil Rehabilitation Law	民事再生法 (12 位 →3 位)	強制執行妨害容疑 (3 位 →25 位) 不動産売却 (9 位 →37 位)
Policy Research Council chairman	政調会長 (10 位 →2 位)	権力 (2 位 →32 位)，派閥 (3 位 →24 位)
Firearms and Swords Control Law	銃刀法 (4 位 →1 位)	元組長 (1 位 →13 位) 群馬県警 (2 位 →14 位)
sarin nerve gas attack	サリン事件 (23 位 →3 位)	オウム真理教 (7 位 →13 位) 松本智津夫被告 (4 位 →18 位)

な範囲に限定することが多い．一方，本稿の実験では，日英関連報道記事から得られる訳語候補 [3] を用い，それらの候補の間の順位付けの性能を評価した．また，文献 [1] では，連語を構成する単語の訳語の組合せを訳語候補として，ウェブ検索エンジンにより収集した二言語非対訳文書を用いて，訳語対応の推定を行っている．この方法では，対訳関係を推定するタームの構成単語の間に訳語の関係がある必要があるが，本稿の手法ではではそのような制限は設けておらず，一般のタームの間の訳語対応推定に適用可能である．

5. おわりに

本稿では，ウェブ検索エンジンを用いて各タームの出現する日英非対訳文書を収集し，これを用いて訳語候補順位付けを行う手法を提案した．評価実験の結果では，報道記事において出現頻度の低いタームについて，ある程度の精度が得られ，ウェブ検索エンジンにより収集される日英非対訳文書の有効性が確認できた．今後は， ウェブ検索エンジンを活用するなどして，報道記事以外の情報源から訳語候補を効率的に収集する技術を確立することが不可欠である．また，訳語候補の順位付けにおいては，競合する候補間でできるだけ異なる文書集合を収集した上で訳語対応推定を行う手法を実現することにより，より高精度な訳語候補の順位付けが行えると考えている．さらに，訳語候補に対して意味分類や固有表現クラスの推定を行ない，これを二言語間で対応付けるというアプローチにより，より高精度な訳語候補の順位付けを実現することが有望と考えている．

文　献

[1] Y. Cao and H. Li. Base noun phrase translation using Web data and the EM algorithm. In *Proc. 19th COLING*, pp. 127–133, 2002.

[2] 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.

[3] 日野浩平, 宇津呂武仁, 中川聖一. 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, Vol. 2004, No. (2004–NL–162), 2004.

[4] 日野浩平, 宇津呂武仁, 中川聖一. 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, 2004.

[5] 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).

[6] 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.