ウェブ上の日英非対訳文書を用いた訳語対応推定*

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 †† 豊橋技術科学大学 工学部 情報工学系

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 [Utsuro03, 日野 04]．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．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おり，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．

*Estimating Bilingual Term Correspondences from Japanese-English Non-parallel Documents collected from WWW

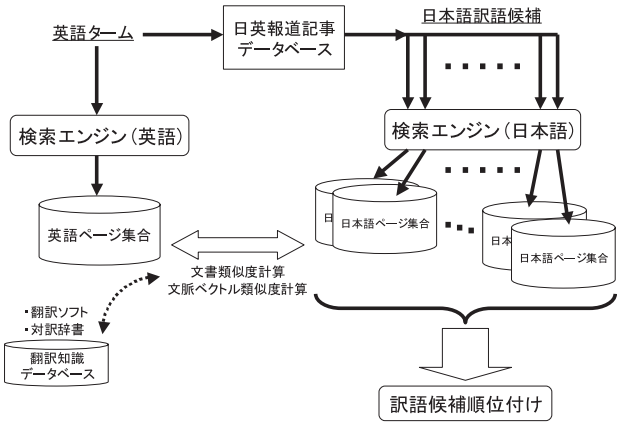

図 1: ウェブ検索エンジンを利用した訳語推定の流れ

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 [日野 04] 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．

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

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[1]．

1. $0 < h(t_E) \leq 100$ の場合，
   $h_L = 0$, $h_U = 10,000 \times h(t_E)$
2. $100 < h(t_E) \leq 20,000$ 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．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」[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 [日野 04] と同等のサイズ[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．

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

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，[日野 04] で選定した英語ターム 100 個のセット，および，別途選定した 29 個の英語タームセットを用いる[4]．本稿では，これらの各英語タームにつき，[日野 04] の英日方向の 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．

まず，29 英語タームについて，検索エンジンによって収集できる最大数 (今回用いた検索エンジンでは 1,000 件) の文書を収集した．今回の実験では，ヒット数が 1,000 未満のタームが含まれていた．また，html ファイルの形式上有用なテキストを含まないと判定したページは削除した．その結果，一タームあたりの平均文書数は，英語で 495.7，日本語で 680.8 となった．また，2.4.2 節の文書間類似度を用いた訳語対応推定では，これらの最大 1000 文書のうちで，各タームにつき最大 100 文書だけを用いて訳語対応の推定を行った．

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．3.4 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

[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．

[2] `http://chasen.aist-nara.ac.jp/`

[3] 英語文書 200〜600 単語，日本語文書 1,500〜4,000 バイト．

[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．また，日英関連報道記事を用いる [日野 04] の訳語対応推定の方法により，[日野 04] で示した性能とほぼ同等の性能で訳語候補の順位付けが行える．

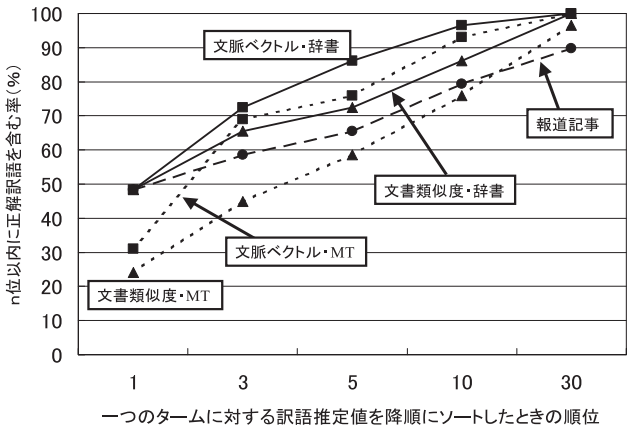

図 2: 翻訳ソフトと対訳辞書の精度比較 (29 ターム, 文書類似度の閾値 $L_d = 0.2$)

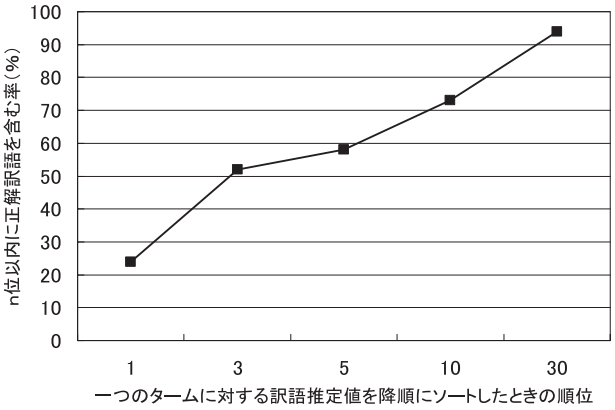

図 3: 対訳辞書・文脈ベクトルを用いた訳語候補順位付けの性能 (100 ターム)

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．

3.2 訳語対応推定精度の評価

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 2 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 (今回の実験では，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．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．また，図 2 には，比較のために，約三年分の日英報道記事から推定した英日方向の ϕ^2 統計値による訳語候補順位付け [日野 04] の性能も示す．

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 [日野 04] 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．

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．

次に，100 英語タームを対象として，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結果を図 3 に示す．英語 29 タームを対象とした図 2 の結果における対訳辞書・文脈ベクトルの組の性能と比較すると，全体的に性能が下がっていることが分かる．これにはいくつかの要因が考えられるが，現時点では，29 タームの方が，100 タームよりも，訳語候補の順位付けが容易であるタームセットとなっていたと考えている[5]．

[5] 100 タームを対象とした実験においては，日本語検索エンジンを利用する際の不手際があり，29 タームを対象とした図 2 の実験と同一条件とはなっていない．現在，図 2 の実験と同一条件での再実験を行っており，図 3 よりも精度が改善する可能性がある．しかし，100 タームを対象とした実験と同一の条件で，29 タームに対する再実験を行った結果では，図 2 の結果よりも 7%程度精度が低下しただけであった．したがって，100 タームを対象とした再実験における精度の改善は，極端に大きくないことが予想される．

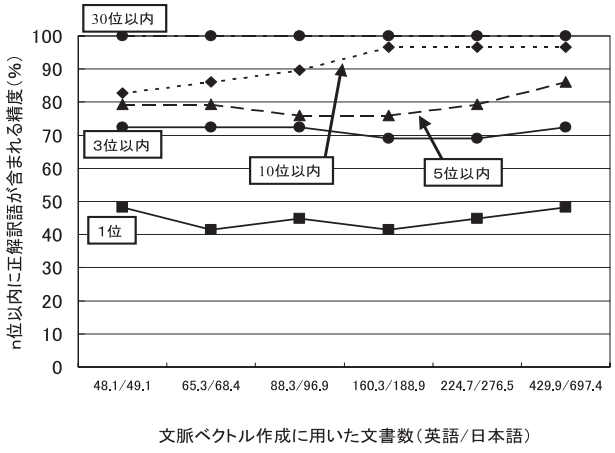

図 4: 文書数と訳語対応推定精度の相関 (29 ターム)

3.3 報道記事を用いた訳語対応推定との比較

図 2 に示すように，29 タームを評価用英語タームとした実験においては，ウェブ検索エンジンにより収集した文書を用いた訳語候補順位付けの性能は，日英報道記事を用いた訳語候補順位付け [日野 04] の性能とほぼ同等である．一方，100 英語タームを評価対象とした場合の訳語候補順位付けの性能は，現時点では図 3 のようになっている．日英報道記事を用いた訳語候補順位付け [日野 04] においては，この 100 タームと 29 タームの間に大きな精度差はないことが分かっている．したがって，現時点では，ウェブ検索エンジンにより収集した日英非対訳文書を情報源として訳語候補順位付けを行った結果においては，報道記事を用いた訳語候補順位付けよりも精度が低下する可能性が高いと言える．しかし，図 3 においても，ある程度の精度は保っており，ウェブ検索エンジンにより収集される日英非対訳文書は，訳語候補順位付けのタスクにおいて有用な情報源であると言える．

3.4 文書数と訳語対応推定精度の相関

29 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 4 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，最大文書数が 200 から 100 になるところで，上位 10 以内までの精度が下がり始める以外には，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 100 文書程度を収集すればほぼ十分であると言える．

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，[Fung98]) 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 [日野 04] 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，[Cao02] 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おり，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．

参考文献

[Cao02] Cao, Y. and Li, H.: Base Noun Phrase Translation Using Web Data and the EM Algorithm, *Proc. 19th COLING*, pp. 127–133 (2002).

[Fung98] Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, *Proc. 17th COLING and 36th ACL*, pp. 414–420 (1998).

[日野 04] 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 (2004).

[Utsuro03] Utsuro, T., Horiuchi, T., Hamamoto, T., Hino, K. and Nakayama, T.: Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora, *Proc. 10th EACL*, pp. 355–362 (2003).