ウェブからの関連語収集手法を用いた専門用語の訳語推定*

日野 浩平[†]　佐々木 靖弘[††]　宇津呂 武仁[††] 土屋 雅稔[†††] 中川 聖一[†] 佐藤 理史[††]
[†]豊橋技術科学大学 工学部 情報工学系　[††]京都大学 情報学研究科
[†††]豊橋技術科学大学 情報処理センター

1　はじめに

近年，ウェブ上のサイトにおいては，多種多様な専門分野の日本語・英語文書が存在する。これらの文書においては，日々最新の情報が公開されており，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．一方，従来の訳語対獲得においては，対訳コーパスやコンパラブルコーパスなどの多言語コーパスから，様々な獲得手法の研究が行なわれてきた [Matsumoto00]．これらの研究では，コーパス中に出現するものを訳語対の候補としていた．例えば，我々はこれまでに，与えられた報道記事コーパス中に複数回出現している訳語対を獲得する手法を提案した [Utsuro04]．報道記事では，内容の関連した記事が日本語記事，英語記事中に出現しているため，対訳コーパスから訳語対を獲得する手法を適用することにより，比較的容易に訳語対が獲得できる．しかし，実際には報道記事中に出現しない多くのタームが存在し，そういった訳語は獲得不可能である．

そこで，本稿では，実際に訳語を知りたいタームを与え，より広範な情報源であるウェブ上からコーパスを自動収集し．訳語の獲得を行なう手法を提案する．

2　概要

本稿の手法では，関連語を利用して同一分野の二言語の文書を収集し，これをコンパラブルコーパスとみなして翻訳知識の獲得を行なう．関連語を用いてコンパラブルコーパスを収集する手法の基本的考え方を以下に示す．まず，日本語と英語の文書が類似しているなら，それぞれの文書に現れる語彙も類似するはずである．例えば “natural language” と “morphological analysis” が出現する英語の文書と「自然言語」と「形態素解析」が出現する日本語の文書はお互いに内容が似ていると予想できる．ここで，日本語文書中に出現する訳語が未知のターム t_J が「自然言語」と「形態素解析」と共起するなら，同様に t_J の訳語 t_E も “natural language” および “morphological analysis” と共起している可能性が高い．この考えを利用し，訳語が未知である語とよく共起する語 (ここでは関連語と呼ぶ)，およびその関連語の訳語を利用して，内容的に類似した日本語文書，英語文書のコンパラブルコーパスを収集する．

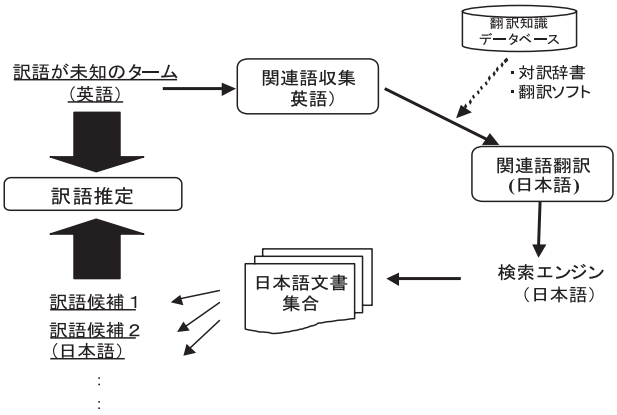

図 1: 関連語収集手法を用いた訳語推定

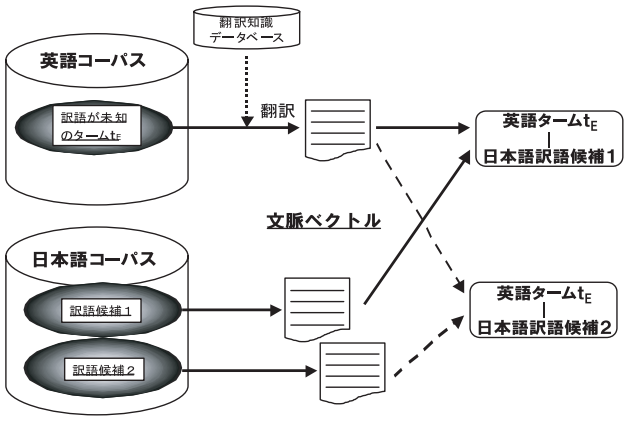

図 2: 文脈ベクトルに基づく訳語対応の推定

ウェブからの関連語収集手法を用いた専門用語の訳語推定の流れを図 1 と図 2 に示す．本稿では，便宜上，英語ターム t_E を与え日本語訳語を得る手法を説明するが，本稿で提案する手法は，英語から日本語への方向に限定されるものではない．以下，図 1 について説明する．まず，訳語を推定したい英語ターム t_E を与える．この英語ターム t_E に対してウェブ上の検索エンジンを利用して，英語関連語を収集する．そして，収集

*Estimating Bilingual Correspondences of Technical Terms through Related Term Collection from Web

された英語関連語と，その英語関連語の日本語訳語を用いて，日本語の訳語候補を得るための日本語コーパスを収集する．次に，図 2 に示すように，英語ターム t_E とその英語関連語を使って英語コーパスを収集し，英語コーパス中の英語ターム t_E が出現する文を日本語に翻訳し，その結果から名詞・動詞などを要素に持つ文脈ベクトルを作成する．同様に，収集した日本語コーパス中に出現する各日本語訳語候補の文脈に共起する名詞・動詞などを要素に持つ文脈ベクトルを作成する．作成した各文脈ベクトルの距離を，余弦類似度で計り，降順にソートし訳語候補の順位付けを行う．

3 関連語収集

本節では，ウェブ検索エンジンを用いた関連語収集について説明する．ここでは，[佐々木 04] で作成された日本語関連語収集システムの英語版を用いる．これはウェブの検索エンジンを利用して，入力されたターム t_E に関連する用語集合を収集するシステムである．

3.1 コーパス作成

入力された英語ターム t_E に対して，(1)t_E，(2)“what’s AND t_E”,(3)“glossary AND t_E” という 3 種類のクエリを (英語対応の) 検索エンジンに入力し，得られた URL のそれぞれ上位 100 ページを入手する．それぞれのページを整形して文に分割し，ターム t_E を含む文とその前後 50 文を抽出し，コーパスを作成する．

3.2 候補語抽出

次に，作成したコーパスからターム t_E の関連用語の候補となる語を抽出する．コーパス中に出現する英語単語列を Charniak parser[1] で構文解析し，次の品詞列パターンを満たすもののうち，2 回以上出現する複合語を候補とした．ただし，$*$ は 0 回以上の繰り返し，$+$ は 1 回以上の繰り返しを表す．

- W_1 = [形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞] $*$ 名詞
- W_2 = ([形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞]+,) $*$ [形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞] + *and* [形容詞 | 名詞 | 現在分詞 | 過去分詞 | 動名詞] $*$ 名詞

3.3 関連度の計算

ある候補語を y，$H(t_E \vee y)$ を t_E と y の OR 検索のヒット数，$H(t_E \wedge y)$ を t_E と y の AND 検索のヒット数として，t_E と y の間で，以下で定義される関連度を計算する．候補語のうち，関連度の上位のものを英語ターム t_E の関連語とする．

$$R_{\wedge/\vee}(t_E, y) = \frac{H(t_E \wedge y)}{H(t_E \vee y)} \quad (1)$$

4 コンパラブルコーパスの収集

本節では，ウェブ上の文書からコンパラブルコーパスを収集する方法について述べる．

4.1 複合語の構成単語の利用可能性

本節の手法では，英語ターム t_E の関連語や t_E の構成単語と，これらの訳語を利用する．t_E の構成単語とは，t_E が 2 語以上からなる複合語であった場合に，その内部に含まれる単語のことを指す．例えば”mad cow disease”(狂牛病) は”mad”(狂った),”cow”(牛),”disease”(病) の 3 つの構成単語から成る．このように，複合語 t の訳語の構成要素として，t の構成単語の訳を含む場合がある．コンパラブルコーパス収集において，複合語の構成単語を利用する方法の有効性を評価するために，予備調査として，英語複合語の構成単語の訳語が汎用対訳辞書 (ここでは，英辞郎 Ver.79 (129 万語)) に含まれている割合を調べた．まず，コンピュータ用語辞典 (英和 33,000 語)[2] ，および，医学用語辞典から，英語複合語の専門用語を抜き出した[3] ．この複合語の構成単語を汎用対訳辞書で翻訳し，構成単語の日本語訳が複合語の日本語訳に含まれている割合を調査した (表 1)．複合語のうち，構成単語訳が一つ以上含まれているものは 7 割以上存在した．これより，複合語の構成単語の訳語も，コーパス収集の重要な手がかりとなると言える．

4.2 コンパラブルコーパスの収集

英語ターム t_E の関連語・構成単語，および，ウェブ検索エンジンを利用して，英語コーパスを収集する方法，および，日本語コーパスを収集する方法を以下に述べる．まず，英語ターム t_E の m 個の構成単語を $I_E^1, I_E^2, \cdots, I_E^m$，その訳語を $I_J^1, I_J^2, \cdots, I_J^m$ とする．また英語ターム t_E の関連語を $C_E^1, C_E^2, \cdots, C_E^n$ とし，その訳語を $C_J^1, C_J^2 \cdots C_J^n$ とする．

4.2.1 日本語コーパスの収集

日本語に対応した検索エンジンから，文書を収集する．英語関連語のうち，汎用対訳辞書 (英辞郎，Ver.79 (129 万語)) のエントリに含まれる英語連語のみを，関連語の候補として関連度の計算を行ない，その上位三つの訳語 C_J^1, C_J^2, C_J^3 で，ウェブ検索エンジンの AND 検索ヒット数を調べる．ここでヒット数が下限値の 20 を越えな

[1] http://www.cs.brown.edu/people/ec/

[2] 日外アソシエーツ株式会社 第 3 版.

[3] 医学用語辞典については専門用語一定数を無作為抽出.

表 1: 複合語専門用語の構成単語の訳語が汎用対訳辞書に含まれる割合

	複合語専門用語辞書		
	A		B
	個数	割合 (%)	割合 (%)
全ての構成単語の訳語が含まれる	2092	17	15
一部の構成単語の訳語が含まれる	8652	70	59
含まれない	1698	13	26
計	12442	100	100

複合語専門用語辞書 A: コンピュータ用語辞典 (3 万 3 千語)
複合語専門用語辞書 B: 医学用語辞典
汎用対訳辞書 英辞郎 Ver.79 (129 万語)

い場合は，AND 検索に用いる関連語の数を関連度の上位二つに減らす．また，汎用対訳辞書のエントリに，日本語訳候補が二つ以上ある場合，ウェブ検索エンジンの AND 検索ヒット数において最大のヒット数を持つ訳の組合せを選択する．この関連語の訳語の AND 検索により得られた上位 100 ページを文書集合 $Text(C_J) = Text(C_J^1 \wedge C_J^2 \wedge C_J^3)$ (あるいは，$Text(C_J^1 \wedge C_J^2)$) とする．同様に，英語構成単語 $C_E^1, C_E^2, \cdots, C_E^m$ の日本語訳候補の組合せのうち，ウェブ検索エンジンの AND 検索ヒット数が最大となる組合せ $I_J^1, I_J^2, \cdots, I_J^m$ を用いたクエリにより，検索エンジンから得られた上位 100 ページを文書集合 $Text(I_J) = Text(I_J^1 \wedge I_J^2 \wedge \cdots \wedge I_J^m)$ とする．そして，これらの文書集合の和集合 $Text(C_J) \cup Text(I_J)$ を日本語コーパスとする．(実際の検索エンジン上では $\wedge$ は AND である．)

4.2.2 英語コーパスの収集

日本語コーパス収集に用いた C_J^1, C_J^2, C_J^3 (または C_J^1, C_J^2) に対応する英語関連語 C_E^1, C_E^2, C_E^3(または C_E^1, C_E^2)，および，英語ターム t_E を用いてクエリを作成し，英語に対応した検索エンジンを用いて上位 100 ページを収集する．そして，英語ターム t_E を含む文の前後 1 文ずつの合計 3 文を抜き出した英語文書集合 $Text(t_E \wedge C_E^1 \wedge C_E^2 \wedge C_E^3)$ (または $Text(t_E \wedge C_E^1 \wedge C_E^2)$) を英語コーパスとする．

5 文脈ベクトルを用いた訳語対応の推定

5.1 日本語訳語候補の収集

日本語形態素解析システム「茶筌」[4] を用いて，収集した日本語コーパスに出現するすべての語に品詞を付与し，記号などの不要語を除く接頭詞，名詞，未知語によって構成される任意の列を，日本語訳語候補とする．ただし，計算機の処理上の都合から，日本語訳語候補は，単語数 5 以下の複合語・単語に限定した．

5.2 文書の頻度ベクトル化

次に，4.2 節において収集した html 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.79，129 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を「茶筌」により形態素解析し，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．

5.3 訳語対応推定

文脈ベクトルの類似性を用いて訳語対応推定を行う．英語ターム t_E および日本語ターム 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．

6 実験および評価

6.1 評価用訳語対

4.1 節で予備調査に用いたコンピュータ用語辞典に含まれる訳語対のうち，既存の対訳辞書 (英辞郎， Ver.79，129 万語) に含まれず，かつ，5.2 節の文書翻訳において用いた翻訳ソフトで訳せないものを抜き出した．さらに，訳語対の英語タームおよび日本語タームに対して，検索エンジンでの検索ヒット数を調べ，ページヒット数 100 以上の訳語組 100 組を無作為に選定し，評価用訳語対とした．

6.2 関連語収集の評価

関連語収集によって得られた英語関連語の訳語 C_J^1, C_J^2, C_J^3(または C_J^1, C_J^2) が，英語ターム t_E の正解日本語訳語 t_J と共起している割合を調べるため，ウェブ

[4] http://chasen.aist-nara.ac.jp/hiki/ChaSen/

表 2: 英語関連語の日本語訳語と正解日本語タームとの AND 検索ヒット数

	割合 (%)
ヒット数 10 以上	68
ヒット数 20 以上	58
ヒット数 50 以上	46
ヒット数 100 以上	34

表 3: 各文書集合のサイズと日本語訳語候補数

	平均文書サイズ (byte)	平均訳語候補数
英語	13,143	-
日本語	985,134	7,920

検索エンジンに $C_J^1 \wedge C_J^2 \wedge C_J^3 \wedge t_J$ (または $C_J^1 \wedge C_J^2 \wedge t_J$) のクエリを与え，検索ヒット数を調査した．この調査を評価用訳語組 100 組に対して行なった結果を表 2 に示す．C_J^1, C_J^2, C_J^3(または C_J^1, C_J^2) が正解日本語訳語 t_J と関連しているなら，共起 (AND 検索) ヒット数は大きくなり，実際に獲得できる可能性も高くなる．逆にヒット数が小さいなら，共起している割合は低く獲得は困難であると言える．

6.3 訳語対応推定精度の評価

評価用訳語対 100 対に対して，4.2.1 節で収集した日本語コーパス，および，4.2.2 節で収集した英語コーパスのサイズを平均したもの，および，日本語コーパスから収集した日本語訳語候補の平均数を表 3 に示す．英語コーパスのサイズが日本語コーパスのサイズと比較して小さいのは，英語ターム t_E 周辺の文脈しか収集しないためである．

次に，文脈ベクトルを用いた訳語対応推定精度において，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方法を比較した結果を図 3 に示す．ここでは，翻訳ソフトを用いた場合の方が高い性能を示した．翻訳ソフトを用いた場合，1 位で推定できた訳語対は全体の 25% 近くであった．正解日本語訳語の順位が一位となった例を表 4 に示す．また，およそ半分近くの訳語対において，100 位以内の訳語候補の中に正

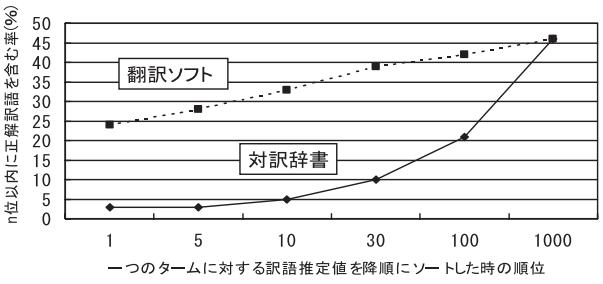

図 3: 訳語推定性能：翻訳ソフトと対訳辞書の比較

表 4: 正解日本語訳語の順位が一位となる例

英語ターム	日本語訳語
diagonal method	対角線論法
address translation table	アドレス変換テーブル
initial-value problem	初期値問題

解日本語訳語を含んでいる．なお，全訳語候補中に正解日本語訳語が含まれる場合は，ほとんどの場合において，100 位以内に正解日本語訳語を含んでいた．

7 関連研究

従来，コーパスを用いた訳語対応推定の研究においては，限られたコーパスの中で翻訳知識を獲得する手法の研究が多く行なわれてきた．しかし，これらの研究では，コーパス中に出現しない語彙に対しての訳語の獲得は不可能である．これに対して，最近の研究における特徴としては，ウェブ上のテキストを利用してコーパスを収集し，そのコーパスから訳語を獲得するといった点が挙げられる．例えば，[Cheng04] では，英語タームを検索質問として，ウェブ上の中国語ページを収集した結果から中国語訳語候補を生成し，中国語訳語候補と英語タームとの間の統計的相関，および，文脈ベクトルの類似性を併用して，英語・中国語間の訳語対応を推定する手法を提案している．

8 おわりに

本稿では，ウェブからの関連語収集手法を用いた専門用語の訳語推定の手法について述べ，実際に，半分近くの評価用訳語対において，訳語候補の中に正解日本語訳語を含むことを示した．今後は，要素合成法を用いた訳語推定法 [外池 05] との統合を行なう．

参考文献

[Cheng04] Cheng, P.-J., Lu, W.-H., Teng, J.-W. and Chien, L.-F.: Creating Multilingual Translation Lexicons with Regional Variations Using Web Corpora, *Proceedings of the 42nd ACL*, pp. 534–541 (2004).

[Matsumoto00] Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, Dale, R., Moisl, H. and Somers, H. (eds.), *Handbook of Natural Language Processing*, chapter 24, pp. 563–610, Marcel Dekker Inc. (2000).

[佐々木 04] 佐々木靖弘, 佐藤理史, 宇津呂武仁：用語間の関連度を測る指標の提案, 言語処理学会第 10 回年次大会論文集, pp. 25–28 (2004).

[外池 05] 外池昌嗣, 木田充洋, 高木俊宏, 宇津呂武仁, 佐藤理史：要素合成法を用いた専門用語の訳語候補生成・検証, 言語処理学会第 11 回年次大会論文集 (2005).

[Utsuro04] Utsuro, T., Hino, K., Kida, M., Nakagawa, S. and Sato, S.: Integrating Cross-Lingually Relevant News Articles and Monolingual Web Documents in Bilingual Lexicon Acquisition, *Proceedings of the 20th COLING*, pp. 1036–1042 (2004).