ウェブと要素合成法を用いた専門用語訳語推定

外池　昌嗣[†]　　宇津呂　武仁[†]　　佐藤　理史[††]

[†]京都大学 情報学研究科　[††]名古屋大学 工学研究科

1. はじめに

本論文では，専門用語の対訳集の生成に関する問題を研究する．専門用語の対訳集の生成には，訳語推定の技術が必要不可欠である．これまでに行われてきた訳語推定の方法の１つに，パラレルコーパスまたはコンパラブルコーパスを用いる方法がある．これらの手法では，原言語の用語とその訳語の間の文脈的な類似性を言語を横断して測定し，その文脈的類似性に従ってすべての訳語候補を順位付けする．しかしながら，訳語対応の推定の目的に利用できるパラレルコーパスやコンパラブルコーパスは限られている．それゆえ，それらの既存手法を専門用語の訳語推定のタスクに適用しようとしても，多くの場合，訳語推定対象の専門分野の既存コーパスを見つけることは極めて難しい．

一方，単言語コーパスを用いて要素合成法による訳語推定を行う技術 [1),2)] はより実用的である．なぜならば，単言語コーパスはパラレルコーパスやコンパラブルコーパスに比べて収集が容易だからである．この技術では，ある用語の訳語候補は，その用語の構成要素の訳語を結合することによって構成的に生成される．その後，生成された訳語候補は，専門分野コーパスを用いて検証される．

要素合成法による訳語推定法の有効性を調査するために，既存の専門用語対訳辞書の 10 分野から，日本語と英語の専門用語で構成される訳語対を 667 個無作為に抽出した．そして，それぞれの訳語対が構成的であるかを調べたところ，88%が実際に構成的であるという結果が得られた．このことから，専門用語に対して要素合成法による訳語推定法を適用することは有効である可能性が高いことがわかった．

しかしながら，たとえ単言語であっても，専門分野コーパスを収集するのは高価である．そこで，様々な分野の文書が利用可能なウェブを利用する方法を採用する．ウェブを訳語候補の検証に利用する場合，サーチエンジンを通してウェブ全体を用いて訳語候補の検証を行う方法と [3)]，訳語推定の前にあらかじめ，ウェブから専門分野コーパスを収集しておき，その後，そのコーパスを用いて訳語候補の検証を行う方法 [4)] の２つの方法が考えられる．前者はカバレージに優れる一方，後者は特定分野の大量の用語の訳語推定を行う場合効率がよい．本論文では，要素合成法による訳語推定における２つのウェブ利用法を，カバレージと精度の観点から量的に比較する．

より詳細には，要素合成法による訳語推定において，訳語候補のスコア関数を,「対訳辞書スコア」と「コーパススコア」の２つの要素に分解する．本論文では，それぞれのスコアに対して数種類の実装を設定し，それらを組み合わせることにより，９つのタイプのスコア関数を定義した．実験の結果，コーパスに正解訳語が存在する場合においては，あらかじめウェブからコーパスを収集する方法が，訳語候補の検証時にサーチエンジンを利用する方法より性能がよいことがわかった．さらに，複数のスコア関数が支持する訳語候補が一致する場合のみ，その訳語候補を出力する方法を設計し，評価を行った．実験の結果，単一のスコア関数の精度よりもはるかに高い精度を達成できることがわかった．

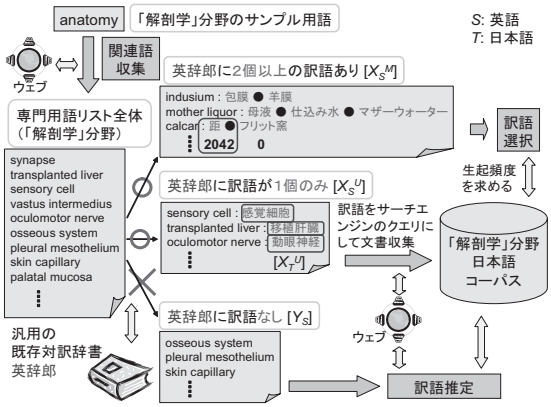

図 1　ウェブを用いた専門用語対訳集生成

2. 専門用語対訳集生成の全体像

ウェブを利用した専門用語対訳集生成の全体像を図 1 に示す．ある特定の分野の用語のサンプルが与えられていると仮定し，対訳集の見出し語にすべき専門用語を，ウェブからの関連語収集手法 [5)] を用いて収集する．それらの収集された専門用語は，既存対訳辞書に掲載されている訳語の個数にしたがって，３つの部分集合に分けられる．すなわち，訳語の数が１である用語の集合 X_S^U，訳語の個数が２以上である用語の集合 X_S^M，訳語が掲載されていない用語の集合 Y_S の３つである．X_S^M の用語に対しては，既存の対訳辞書にある訳語の中から最も適切なものを選択する必要がある．例えば，解剖学分野に属する英語の専門用語 “calcar” の訳語としては，窯業用語の「フリット窯」ではなく，用語「距」が選択されなければならない．一方，Y_S の用語に対しては，訳語候補の生成と検証が求められる．本論文では，上記の２つのタスクのうち，後者のウェブを用いた訳語候補の生成と検証に焦点をあてる．前章で述べたように，ここでは，訳語候補の検証のための２つのウェブ利用法を実験的に比較する．１つ目の方法はサーチエンジンを用いる一方，２つ目は

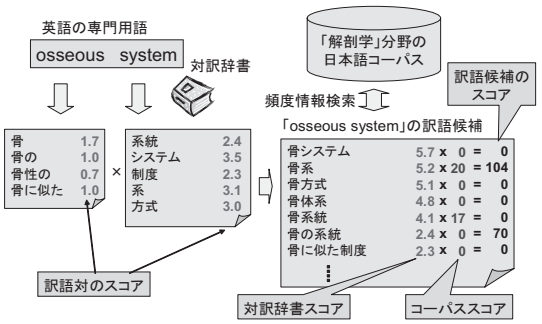

図 2 日本語の専門用語「応用行動分析」の要素合成法による訳語推定

あらかじめウェブから収集した専門分野コーパスを用いる．ここで，既存の辞書に訳語が 1 個だけ掲載されている X_S^U の用語の訳語の集合を X_T^U とすると，2つ目のアプローチでは，X_T^U の用語をサーチエンジンのクエリーとして用いてウェブから専門分野コーパスを収集する．[4)]

3. 要素合成法による専門用語訳語推定

3.1 訳語推定の概要

要素合成法による訳語推定の例として,「解剖学」分野の英語の専門用語 “osseous system” の日本語訳を推定する様子を図 2 に示す．まず，専門用語 “osseous system” を構成要素 “osseous” と “system” に分割し，それぞれの構成要素を既存の対訳辞書を利用して目的言語に翻訳する．このとき，それぞれの構成要素の訳語対にはスコアが付与されている．次に，前置詞句の構成を考慮した語順の規則にしたがって，それらの構成要素の訳語を結合し，訳語候補を生成する．ここで，訳語候補のスコアは，対訳辞書スコアとコーパススコアの積で計算される．対訳辞書スコアは訳語対のスコアの積で計算され，コーパススコアは「解剖学」分野の日本語コーパスから得られる頻度情報に基づいて計算される．最後に，スコア 1 位の訳語候補が選択される．

本研究では，既存の対訳辞書として,「英辞郎」☆に加えて，英辞郎の訳語対から作成した部分対応対訳辞書 [1),4)] を用いる．まず，既存の対訳辞書から，日本語及び英語の用語がそれぞれ 2 つの構成要素からなる訳語対を抽出し，これを別の対訳辞書 P_2 とする．次に，P_2 中の訳語対の第一構成要素から前方一致部分対応対訳辞書 B_P を作成し，第二構成要素から後方一致部分対応対訳辞書 B_S を作成する．さらに，B_P と B_S を併合し部分対応対訳辞書 B を作成する．

3.2 訳語候補のスコア

この節では，要素合成法による訳語推定における訳語候補のスコアを定義する．まず，y_S を訳語推定すべき専門用語とし，y_S は以下に示すように構成要素に分解されると仮定する．

$$y_S = s_1, s_2, \cdots, s_n \tag{1}$$

☆ http://www.eijiro.jp/

ここで，s_i は語の列を表す．次に，y_S に対して生成された訳語候補を y_T と記述するものとすると，y_S は以下のように表すことができる．

$$y_T = t_1, t_2, \cdots, t_n \tag{2}$$

ここで，t_i は s_i の訳語を表す．このとき，訳語対 $\langle y_S, y_T \rangle$ は以下のように表される．

$$\langle y_S, y_T \rangle = \langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, \cdots, \langle s_n, t_n \rangle \tag{3}$$

生成された訳語候補のスコアは，以下に示すように，対訳辞書スコア $Q_{dict}(y_S, y_T)$ とコーパススコア $Q_{coprus}(y_T)$ の積で定義される．

$$Q(y_S, y_T) = Q_{dict}(y_S, y_T) \cdot Q_{coprus}(y_T) \tag{4}$$

対訳辞書スコア $Q_{dict}(y_S, y_T)$ は y_S と y_T の対応の適切さを測定する．コーパススコア $Q_{coprus}(y_T)$ は，訳語候補 y_T の適切さを目的言語コーパスに基づいて測定する．ここで，ある訳語候補が 2 つ以上の系列の訳語対から生成される場合，その訳語候補のスコアはそれぞれの系列のスコアの和で定義される．

対訳辞書スコア

本論文では、2 種類の対訳辞書スコアを導入して比較を行う．2 つのスコアは，以下に示すように，前節で示した対訳辞書に含まれる訳語対のスコアの積で定義される．

- 頻度-構成要素数

$$Q_{dict}(y_S, y_T) = \prod_{i=1}^{n} q(\langle s_i, t_i \rangle) \tag{5}$$

このスコアは，訳語対の構成要素数と，部分対応対訳辞書 B_P, B_S の訳語対の頻度に基づく．このスコアでは、対訳辞書の訳語対に対して，文献 [4)] で提案された優先順位を仮定する．

訳語対 $\langle s, t \rangle$ のスコア $q(\langle s, t \rangle)$ の定義として，$\langle s, t \rangle$ がどの対訳辞書に出現するかで場合分けを行った以下の式を採用する．

$$q(\langle s, t \rangle) = \begin{cases} 10^{(compo(s)-1)} & \text{英辞郎の場合} \\ \log_{10} f_p(\langle s, t \rangle) & B_P\text{の場合} \\ \log_{10} f_s(\langle s, t \rangle) & B_S\text{の場合} \end{cases} \tag{6}$$

ここで，$compo(s)$ は s の構成要素数を表すものとし，$f_p(\langle s, t \rangle)$ は，対訳辞書 P_2 中に第一要素として $\langle s, t \rangle$ が出現する回数を表すものとし，$f_p(\langle s, t \rangle)$ は，P_2 中に第二要素として $\langle s, t \rangle$ が出現する回数を表すものとする．

- 確率

$$Q_{dict}(y_S, y_T) = \prod_{i=1}^{n} P(s_i|t_i) \tag{7}$$

このスコアは，条件付き確率 $P(s_i|t_i)$ の積で定義される．$P(s_i|t_i)$ は対訳辞書及び部分対応対訳辞書を用いて計算される．$P(s|t)$ の定義を以下に示す．

$$P(s|t) = \frac{f_{prob}(\langle s, t \rangle)}{\sum_{s_j} f_{prob}(\langle s_j, t \rangle)} \tag{8}$$

式 (8) 中の $f_{prob}(\langle s, t \rangle)$ は，$\langle s, t \rangle$ が対訳辞書に生起する頻

表 1 9つの訳語候補のスコア関数と，それらの構成要素

スコア ID	対訳辞書スコア		コーパススコア			コーパス	
	頻度-構成要素数	確率	確率	頻度	生起の有無	off-line	on-line (サーチエンジン)
A		prune/final	prune/final			o	
B		prune/final		prune/final		o	
C	prune/final		prune/final			o	
D	prune/final				prune	o	
E	prune/final						
F	prune/final			final	prune	o	
G	prune/final			prune/final		o	
H	prune/final			final		o	
I	prune/final			final			o

度を表す．ただし，$\langle s,t\rangle$ が英辞郎に出現する場合は，頻度が 10 であるとみなし，部分対応対訳辞書 B に出現する場合は，対訳辞書 P_2 に生起する頻度を表す．☆ $f_{prob}(\langle s,t\rangle)$ の定義を以下に示す．

$$f_{prob}(\langle s,t\rangle) = \begin{cases} 10 & \text{英辞郎の場合} \\ f_B(\langle s,t\rangle) & B \text{ の場合} \end{cases} \quad (9)$$

コーパススコア

以下に示す３つのタイプのコーパススコアを評価した．

- 確率: 以下のバイグラムモデルによって推定される訳語候補 y_T の生起確率

$$Q_{corpus}(y_T) = P(t_1) \cdot \prod_{i=1}^{n} P(t_{i+1}|t_i) \quad (10)$$

- 頻度: 目的言語コーパスにおける訳語候補 y_T の生起頻度

$$Q_{corpus}(y_T) = freq(y_T) \quad (11)$$

- 生起: 目的言語コーパスに訳語候補 y_T が生起するかどうかに応じて値を定める

$$Q_{corpus}(y_T) = \begin{cases} 1 & \text{生起する場合} \\ 0 & \text{生起しない場合} \end{cases} \quad (12)$$

スコア関数のバリエーション

表 1 に示すように，本論文では，対訳辞書スコアとコーパススコアの９つの組み合せに関して調査を行った．この表において，‘prune’ は，そのスコアが動的計画法のアルゴリズムを用いた訳語候補生成の過程において，生成された訳語候補の部分列の順位付けと枝刈りに用いられることを示す．‘final’ は，そのスコアが生成された訳語候補の最終結果の順位付けに用いられることを示す．列「コーパス」において，‘off-line’ は，あらかじめウェブから専門分野コーパスを収集し，その後，このコーパスを用いて生成された訳語候補の検証を行うことを示す．‘on-line’ は，サーチエンジンを通して直接的に訳語候補の検証を行うことを示す．

ここで，スコア関数 ‘A’ は，文献 [1] によって提案されたモデルに，スコア関数 ‘D’ は，文献 [4] で提案されたモデルに対応する．スコア関数 ‘E’ は，スコア関数 ‘D’ からコーパススコアの要素を取り除いたものである．一方，スコア関数 ‘I’ は，文献 [3] で提案されたサーチエンジンを通して直接的に訳語候補の検証を行う方法の評価を意図している．

☆ 英辞郎の訳語対の頻度の定義が妥当かどうか経験的に調査する必要がある．

表 2 評価用訳語対の数

辞書名	カテゴリ名	$\lvert Y_S\rvert$	$\lvert X_S^U\rvert$	$C(S)$
マグローヒル	電磁気学	33	36	85%
科学技術用語	電気工学	45	34	71%
大辞典	光学	31	42	65%
岩波	プログラム言語	29	37	93%
情報科学辞典	プログラミング	29	29	97%
コンピュータ用語辞書	(コンピュータ)	100	91	51%
	解剖学	100	91	86%
25 万語	疾患	100	91	77%
医学用語	化学物質・薬物	100	94	60%
大辞典	物理化学・統計学	100	88	68%
合計		667	633	72%

$C(S)$:Y_S の用語のうちコーパスに正解訳語が含まれる割合

3.3 ２つのスコア関数の併用

本節では，２つのスコア関数を併用する手法について検討する．２つのスコア関数は前節で導入した９つの関数から選択する．この手法では，まず，２つのスコア関数によって，専門用語の訳語の信頼度を測定する．次に，両方のスコア関数で 1 位となった訳語候補が一致した場合にのみ，この手法は一致した訳語候補を出力する．この手法の目的は，リコールよりも精度を重視することである．

4. 実験と評価

4.1 評価用訳語対集合

実験では，図 1 に示した専門用語対訳集生成のフレームワークのうち，訳語推定の部分の評価を行う．本論文の評価では，対訳集の見出し語とすべき専門用語の収集のプロセス及び訳語選択の評価は行わない．訳語推定の部分の評価を行うために，表 2 に示す既存の日英専門用語対訳辞書の 10 分野から，無作為に訳語対を抽出し集合 X_S^U 及び Y_S を作成した．（ここで，Y_S の用語として，１語または 1 形態素からなるものは除外した．）1 章で述べ

表 3 単一のスコア関数の評価の結果

	Y_S 全体		生成可能		生成可能&存在	
ID	top 1	top 10	top 1	top 10	top 1	top 10
A	43.8%	52.9%	63.8%	77.1%	82.0%	98.5%
B	42.9%	50.7%	62.4%	73.8%	83.8%	99.4%
C	43.0%	58.0%	62.7%	84.5%	75.1%	94.6%
D	43.0%	47.4%	62.7%	69.0%	**85.9%**	94.6%
E	33.9%	57.3%	49.3%	83.4%	51.1%	84.1%
F	40.2%	47.4%	58.5%	69.0%	80.2%	94.6%
G	39.1%	46.8%	57.0%	68.1%	78.1%	93.4%
H	43.8%	57.3%	63.8%	83.4%	73.6%	84.1%
I	**49.8%**	57.3%	72.5%	83.4%	74.8%	84.1%

表 4 2つのスコア関数を併用した場合の結果

コーパス	組み合わせ	精度	リコール	$F_{\beta=1}$
off-line/ on-line	A & I	**88.0%**	27.6%	0.420
	D & I	86.0%	29.5%	0.440
	F & I	85.1%	29.1%	0.434
	H & I	58.7%	37.5%	0.457
off-line/ off-line	A & H	**86.0%**	30.4%	0.450
	F & H	80.6%	33.7%	0.476
	D & H	80.4%	32.7%	0.465
	A & D	79.0%	32.1%	0.456
	A & F	74.6%	33.0%	0.457
	D & F	68.2%	35.7%	0.469

たように，X_T^U（X_S^U の用語の訳語の集合）の用語はウェブから専門分野コーパスを収集する際に利用する．訳語推定の評価は集合 Y_S に対して行う．10 カテゴリのそれぞれに対して，表 2 に X_S^U 及び Y_S のサイズを，また Y_S に対しては，収集した専門分野コーパスに正解訳語を含む割合を示す．以下では，原言語 S を英語，目的言語 T を日本語として行った実験の人手による評価の結果を示す．

4.2 単一スコア関数の評価

表 1 に示すスコア関数 A∼I を単独で評価した結果を表 3 に示す．列「Y_S 全体」には，Y_S 全体に対する結果を示す．列「生成可能」には，Y_S の訳語対のうち，要素合成の操作によって生成可能な訳語対のみに対する結果を示す．列「生成可能&存在」には，正解訳語がコーパスに存在し，かつ，要素合成の操作によって生成可能な訳語対のみに対する結果を示す．列 'top n' には，正解訳語がスコア n 位以内に含まれる割合を示す．

列「Y_S 全体」におけるスコア関数 'D' と 'E' の "top 1" を比較すると，コーパスを用いない場合より，用いた場合の方が性能がよいことがわかる．さらに列「Y_S 全体」より，コーパススコアの計算のために直接ウェブのサーチエンジンを利用するスコア関数 'I' の正解率は，あらかじめ収集した専門分野コーパスを用いるスコア関数よりも高いことがわかる．これは，集合 Y_S 全体に対して，収集した専門分野コーパスに正解訳語が含まれる割合が 72%と，あまり高くないことが原因である．一方，列「生成可能&存在」より，正解訳語がコーパスに存在する場合は，'I' を除くほとんどのスコア関数において，スコア関数 'I' の精度よりも高い精度を達成できることがわかる．この結果は，あらかじめウェブから専門分野コーパスを収集し，その後，このコーパスをを用いて生成された訳語候補の検証を行うという方法の有効性を示している．

4.3 2つのスコア関数を併用する手法の評価

2つのスコア関数を併用する手法の評価の結果を表 4 に示す．この結果は，'off-line' コーパスを用いるスコア関数と 'on-line' コーパスを用いるスコア関数の組み合せが，'off-line' コーパスを用いる 2 つのスコア関数の組み合せよりも高い精度を達成する傾向があることを示す．一方で，この結果は，'off-line' コーパスを用いる 2 つのスコア関数の組み合わせ（'A' と 'H' のペア）でも高い精度を達成できることも示している．ここで，スコア関数 'A' と 'H' は，それぞれ，頻度に基づくスコア関数と確率に基づくスコア関数なので，両者はスコア関数のデザインにおいて全く異なった特徴を持っている．

5. むすび

本論文では，専門用語の対訳集の生成に関する問題を研究した．専門用語の訳語推定において，その分野の既存のコーパスを見つけることは極めて困難である．本論文では、そのような専門用語の分野のコーパスをウェブから収集するアプローチを採用した．専門用語の訳語推定の方法としては，要素合成法に基づく訳語推定法を用いた．本論文は，要素合成法による訳語推定法のスコア関数における要素の組み合わせの量的比較に焦点をあてた．実験を通して，専門分野コーパスが要素合成法による訳語推定法の性能を向上させることを示した．

参考文献

1) 藤井敦, 石川徹也: 技術文書を対象とした言語横断情報検索のための複合語翻訳, 情報処理学会論文誌, Vol. 41, No. 4, pp. 1038–1045 (2000).
2) 田中貴秋, 松尾義博: 対訳関係のないコーパスからの複合名詞対訳の表現の獲得, 電子情報通信学会論文誌, Vol. J84-D-II, No. 12, pp. 2605–2614 (2001).
3) Cao, Y. and Li, H.: Base Noun Phrase Translation Using Web Data and the EM Algorithm, *Proc. 19th COLING*, pp. 127–133 (2002).
4) Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T. and Sato, S.: Effect of Domain-Specific Corpus in Compositional Translation Estimation for Technical Terms, *Proc. 2nd IJCNLP, Companion Volume*, pp. 116–121 (2005).
5) Sato, S. and Sasaki, Y.: Automatic Collection of Related Terms from the Web, *Proc. 41st ACL*, pp. 121–124 (2003).
6) Tonoike, M., Kida, M., Takagi, T., Sasaki, Y., Utsuro, T. and Sato, S.: A Comparative Study on Compositional Translation Estimation using a Domain/Topic-Specific Corpus collected from the Web, *Proc. 2nd Web as Corpus Workshop* (2006).