要素合成法を用いた専門用語の訳語候補生成・検証

外池　昌嗣[†]　木田　充洋[†]　高木　俊宏[††]
宇津呂　武仁[†]　佐藤　理史[†]

[†]京都大学 情報学研究科　[††]京都大学 工学部 電気電子工学科

1. はじめに

我々のプロジェクトでは、国際会議に参加する、その分野に精通していない参加者や通訳者のために、特定分野・トピックに対する対訳用語集を自動生成することを目指している。この背景として、現在、専門用語の訳語情報は、英辞郎などの汎用的な辞書を用いるだけでは、一部の用語しかその訳語情報を知ることができないことが挙げられる。このような専門用語の訳語情報は、様々な分野で求められおり、それぞれを人手で作るとなると、大変なコストがかかる。

従来より、訳語情報の収集の研究は行われてきた[1]が、それらの手法は、コーパスに含まれない用語の訳語を獲得することはできない。一方、特定の専門分野の用語の訳語情報を収集する場合、専門分野のコーパスは一般には存在しないので、まず、それをどのように集めるかが問題となる。さらに、対訳用語集に載せるべき専門用語のリストを作成する方法も問題となる。

本稿では、その一つの実現法として、図 1 に示すように、その専門分野における用語・文書などのサンプルを与えて、その分野の訳語対集合 Y_{ST} を作るという枠組みを設定する[2]。この枠組みにおいては、以下の 2 つの部分を実現する必要がある。1 つは、用語集に載せるべき用語を集める部分で、もう 1 つは、用語集に載せるべき用語のうち、訳語が未知のもの Y_S に対して訳語を推定する部分である。このうち、本稿では、後者の訳語推定部分に焦点をあてる。訳語推定過程のうち、高木ら[2]は対訳用語集に載せるべき用語集合が与えられたとき、その分野の訳語を含むコーパスの収集法を提案している。一方、本稿では、図 2 に示すように、訳を知りたい用語の構成要素の訳語を合成することによって、訳語を推定する手法（要素合成法）を提案し、さらに、コーパスと要素合成法を併用して、より高精度に訳語を推定する方法について提案する。

2. 部分対応対訳辞書の生成

要素合成法で訳語を生成する場合、用語の構成要素の訳語としては、汎用の対訳辞書 (本稿では「英辞郎」Ver.79[3]を用いる) に載っている訳語だけでは不十分である。例えば、“applied behavior analysis”（応用行動分析）の訳語を知りたい場合を考えよう。“applied” を汎用の辞書で引くと「応用の」などは載っているが、“applied” →「応用」という対応は得ることができない。ここで必要なのは、複合語中の構成単語がどのように訳されるのが自然

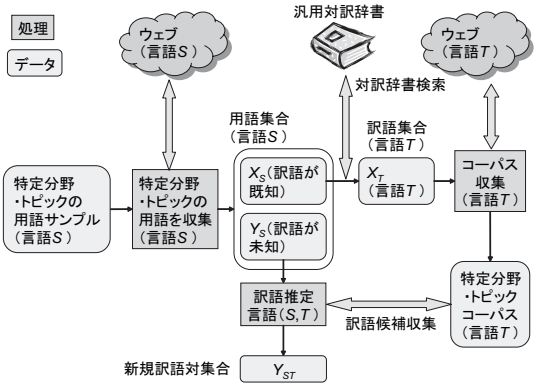

図 1　特定分野・トピックの対訳用語集自動生成

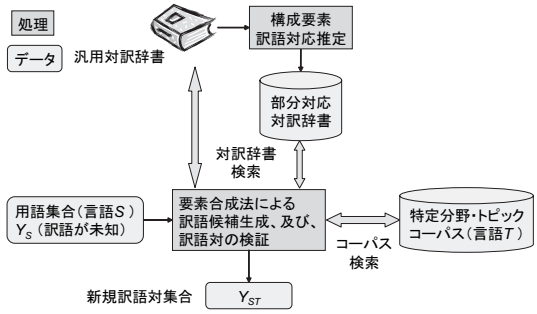

図 2　要素合成法及び特定分野・トピックのコーパスを用いた訳語推定

かという情報である。

このような情報を得るために、汎用の対訳辞書 (英辞郎) から、日英各 2 構成要素からなる訳語対を抽出し (以下、これを辞書 P_2 とする)、そこから日英とも共通の第一要素を持つ訳語対を集め、これらの訳語対の日英双方の第一要素からなる部分対応訳語対を作成する (図 3)。ここで、日本語の構成要素分割には JUMAN[4] を利用した。部分対応訳語対とともに、辞書 P_2 における出現回数も記録しておき、この出現回数を訳語生成時にスコアとして利用する。このようにして作成した部分対応訳語対を集めたものが、前方一致部分対応対訳辞書である。この辞書の訳語対は、訳語候補生成・検証における原言語側の用語の末尾の構成要素以外に適用できるものとする。

同様にして、日英とも共通の第二要素を持つ訳語対を集め、部分対応訳語対を作成し、これらを集めて後方一致部分対応対訳辞書とする。この辞書の訳語対は、原言語側の用語の先頭の構成要素以外に適用できるものとする。

英辞郎及び、作成した 2 つの部分対応訳語辞書の見出し語数及び訳語対数を表 1 に示す。ここで、訳語対とは、2 つの言語の用語を 1 対 1 に対応させたものを指すものとする。英辞郎では、見出し語と訳語の 1 対多の関係が 1 つの単位となっているが、本稿では、この 1 対多の関

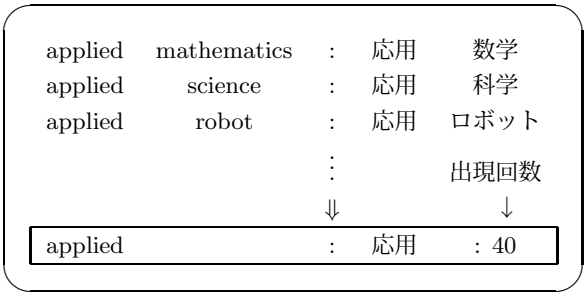

図 3 部分対応訳語対 (前方一致) の作成例

表 1 辞書に含まれる見出し語数及び訳語対数

辞書	見出し語数		訳語対数
	英語	日本語	
英辞郎	1,292,117	1,228,750	1,671,230
辞書 P_2(日英各 2 構成要素)	232,716	200,633	258,211
前方一致部分対応対訳辞書	38,353	38,546	112,586
後方一致部分対応対訳辞書	22,281	20,627	71,429

係を展開し、 1 対 1 の関係としたものを用いる。

3. 要素合成法を用いた訳語候補の生成・検証

3.1 訳語対のスコア

要素合成法により生成される訳語候補にスコアを付与するために、辞書の訳語対にスコアを付与する。利用する複数の辞書の間の優先順序を以下のように定める。

(1) 既存の対訳辞書 (英辞郎) の訳語対で、原言語側が 2 構成要素以上のもの

(2) 部分対応対訳辞書の訳語対で、辞書 P_2 での出現回数が多いもの

(3) 既存の対訳辞書 (英辞郎) の訳語対で、原言語側が 1 構成要素のもの

(4) 部分対応対訳辞書の訳語対で、辞書 P_2 での出現回数が少ないもの

(1) は、構成要素の訳語を組み合わせるよりも、既存の辞書に含まれる訳語対の方が信頼できるという仮定に基づく。(2) は、辞書 P_2 での出現回数が多い部分対応対訳辞書の訳語対は、既存の辞書の 1 構成要素の訳語対よりも信頼できるという仮定に基づく。(4) を最後としたのは、信頼のできない部分対応を排除するためである。

上記を実現するため、訳語対 $\langle s,t\rangle$(ただし、s は原言語側の用語、t は目的言語側の用語) のスコアリング関数 $q(\langle s,t\rangle)$ を以下のように定める。

$$q(\langle s,t\rangle) = \begin{cases} 10^{(compo(s)-1)} & \text{既存の辞書} \\ \log_{10} f(\langle s,t\rangle) & \text{部分対応訳語辞書} \end{cases} \quad (1)$$

ただし、$compo(s)$ は、原言語側の用語 s の構成要素数を表す。また、$f(\langle s,t\rangle)$ は、$\langle s,t\rangle$ 作成時に記録しておいた、$\langle s,t\rangle$ の P_2 中での出現回数を表す。

3.2 動的計画法に基づくアルゴリズム

本稿における要素合成法の実装は動的計画法に基づく。このとき、訳語候補及び、訳語候補の生成過程で作られる部分訳語 y_t が訳語対 $p_1, p_2 \cdots, p_n$ から生成されたとすると、y_t のスコア $Q(y_t)$ は、以下に示すように、構成要素の訳語対 $p_i = \langle s_i, t_i\rangle$ のスコアの積とする。

$$Q(y_t) = \prod_{i=1}^{n} q(p_i) \quad (2)$$

ただし、別の訳語対の組み合わせで同じ訳語候補または部分訳語が生成されたときは、両者のスコアを加算してマージする。実行時には、あらかじめ出力する訳語候補の数 N を決めておき、生成過程においては、スコア上位 N 個に入らない部分訳語を捨てながら生成を進める。

表 2 評価用訳語対集合 (Y_{ST})

辞書名	カテゴリ名	個数	コーパスにおける評価用訳語の収集率	
			日本語	英語
マグローヒル	電磁気学	33	88%	91%
	電気工学	45	76%	84%
	光学	31	87%	81%
岩波 情報科学辞典	プログラム言語	29	93%	100%
	プログラミング	29	93%	97%
コンピュータ用語辞書	(コンピュータ)	100	59%	75%
25 万語 医学用語大辞典	解剖学	100	87%	75%
	疾患	100	85%	76%
	化学物質及び薬物	100	76%	76%
	物理化学及び薬物	100	71%	74%
加重平均		67	78%	79%

3.3 コーパスを利用した訳語候補生成・検証

要素合成法を発展させ、動的計画法の生成過程において、コーパスに存在しない部分訳語が生成された場合、この部分訳語を削除する。この手法は、コーパスによって部分訳語の検証を行いながら訳語候補の生成を行う手法といえる。

4. 実験及び評価

4.1 評価用訳語対

評価実験においては、表 2 に示す、既存の 4 つの日米対訳辞書の 1 0 カテゴリに含まれる訳語対 $\langle s,t\rangle$ のうち、以下に示す条件を満たすものの一部を評価用訳語対集合 Y_{ST} として利用する。

- 英辞郎に s、t とも含まれない
- 日本語側、英語側とも、 2 構成要素以上
- 日本語側の語のヒット数 (goo) が 100 以上かつ、英語側のヒット数 (AltaVista) が 100 以上 10,000 以下

コーパスとしては、文献 [2] で作成したものを用いる。評価用訳語対の個数、及び、コーパスにおける評価用訳語 y_T の収集率 (評価用訳語対集合 Y_{ST} に含まれる評価用訳語 y_T がコーパス中に出現する割合) を表 2 に示す☆。

4.2 要素合成法を用いた訳語候補の生成

上記の評価用訳語対に対して、要素合成法で訳語推定を行った結果を表 3(英→日) に示す☆☆。「正解生成可」と

☆ コーパスにおける評価用訳語対の収集率計算においては、高木ら [2] と違い、出現頻度の下限を 1 としている。この理由は、本稿の場合、要素合成法の過程で生成される部分訳語のうち、コーパスに存在しない部分訳語を削除することが目的だからである。

☆☆ 日英方向の結果の表は省略。ただし、日英方向の訳語推定においては、訳語推定対象の日本語の用語を構成要素に分割する際、JUMAN では「グラフアルゴリズム」の様なカタカナの複合語を分割できな

は、要素合成法と現在の辞書で正解訳語が生成できるかどうかを表し、できるならば○、できないならば×とする。一方、「候補生成可」とは、1つ以上の訳語候補が生成されるかどうかを表す。そして、○/○は正解生成可を、×/○は正解生成不可だが候補生成可を、×/×は候補生成不可を表す。その結果、英日、日英、両方向とも平均19%の割合で1位に正解訳語がランクされた。10位以内に正解訳語がランクされる割合は、英日方向で平均40%、日英方向で平均43%であった。今回用意した辞書資源で正解訳語が生成可能な割合 (○/○) は、両方向とも平均約50%であった。

次に、部分対応訳語辞書の効果を確かめるために、英辞郎を用いない場合と、部分対応辞書を用いない場合について、英日方向で要素合成法による訳語推定の実験を行った。その結果、英辞郎を用いない場合でも、正解生成可の割合は41%とそれほど下がらず、1位が正解の割合は21%とむしろ向上した。一方、部分対応訳語辞書を用いない場合は、正解生成可の割合が27%、1位が正解の割合は8%と、大幅に性能が低下した。

「疾患」のカテゴリにおいて、英日方向の訳語推定で訳語候補が生成できない原因を調べた結果を表4に示す。この表では、何を改善すれば原因が取り除けるかという観点で原因を分類している。訳語が生成できない原因の43%が「操作の不備」に起因するものであり、対処可能であると考えられる。例えば、「of、by、toによる順序交換」であれば、前処理で順序を交換すればよいだろう。一方、「辞書の不備」に起因するものは17%存在する。英辞郎から、日英各3構成要素からなる訳語対を抽出し、これらの訳語対も、部分対応訳語辞書に含めることによって、辞書の拡充を行えば、改善する可能性がある。しかしながら、「改善不可」に分類した25%は要素合成法で訳語を推定するのは困難である。

ここで、評価用訳語対のうち、構成要素が日英で対応していて、要素合成法で生成できる可能性がある訳語対の割合を人手で調べ、このうち、実際に、現在の辞書と要素合成法を用いて、正解訳語を生成できる割合を調べた。日英方向の結果を表5に示す。この結果、まず、評価用訳語対のうち平均88%が要素合成法で生成できる可能性があると判定された。そして、そのうち、平均約57%が現在の辞書と要素合成法で生成可であることがわかった。なお、英日方向でもほぼ同じ結果が得られたが、ハイフンを含む語の扱いの都合で、日英方向よりも平均約1%低い割合となった。

4.3 コーパスを利用した訳語候補生成・検証

次に、コーパスを利用した訳語候補生成・検証の結果を表6(英→日) に示す☆。その結果、コーパスを用いない方法に比べて、1位に正解訳語がランクされる割合が、英日方向で平均10%、日英方向で平均7%向上した。一方、10位以内に正解訳語がランクされる割合は、英日両方では平均3%、日英方向では平均4%低下した。

ここで、コーパスの効果に焦点を絞って評価するために、コーパスに正解が存在し、要素合成法で正解生成可な評価用訳語対のみを対象として、スコア1位の訳語候補が正解となる訳語対の割合を調べた。

$$\text{割合} = \frac{\text{コーパス (未使用/使用) で 1 位が正解}}{\text{コーパスに正解が存在し、正解生成可}} \quad (3)$$

その結果を図4(英→日)、図5(日→英) に示す。コーパスに正解訳語が存在するならば、コーパスを利用するとスコア1位の訳語候補が正解となる割合が著しく向上し、英日方向で平均36%、日英方向で平均26%向上した。した

表 3 要素合成法を用いた訳語候補生成の性能（英→日）

カテゴリ名	正解訳語が n 位以内		正解生成可/候補生成可		
	$n=1$	$n=10$	○/○	×/○	×/×
電磁気学	27%	33%	33%	39%	27%
電気工学	18%	40%	53%	29%	18%
光学	26%	45%	55%	19%	26%
プログラム言語	21%	55%	62%	21%	17%
プログラミング	34%	45%	55%	28%	17%
コンピュータ	23%	51%	59%	32%	9%
解剖学	22%	41%	60%	29%	11%
疾患	10%	26%	34%	50%	16%
化学物質	12%	27%	31%	42%	27%
物理化学	21%	47%	57%	33%	10%
加重平均	19%	40%	49%	35%	16%

表 4 カテゴリ「疾患」の訳語生成の誤り分析 (英→日)

大分類	原因	個数	割合	割合
辞書の不備	辞書にエントリがない	12	16%	17%
	表記のゆれの問題	1	1%	
改善不可	前置詞のない順序交換	5	7%	25%
	非構成的	14	19%	
操作の不備	of, by, to による順序交換	14	19%	43%
	「の」が不要	5	7%	
	「's」の扱い	6	8%	
	「性」、「的」の扱い	3	4%	
	ハイフンの扱い	4	5%	
データの誤り	テストセットの不備	2	3%	4%
	辞書の訳語中の空白	1	1%	
その他	スコアの問題	7	9%	11%
	出力する訳語候補数の問題	1	1%	
合計		75	100%	100%

表 5 人手で構成的と判定された訳語対数および生成可能な訳語対の割合 (日→英)

カテゴリ名	構成的 (人手)		構成的 (人手) のうち、要素合成法＋現在の辞書で生成可
	個数	割合	
電磁気学	27	82%	41%
電気工学	35	78%	71%
光学	29	94%	59%
プログラム言語	27	93%	67%
プログラミング	28	97%	57%
コンピュータ	91	91%	65%
解剖学	91	91%	66%
疾患	86	86%	41%
化学物質	79	79%	42%
物理化学	91	91%	66%
加重平均	67	88%	57%

い。そこで、日本語の構成要素分割では、JUMAN で構成要素分割した結果を人手で修正した。

☆ 日英方向の結果の表は省略

表 6 要素合成法とコーパスを併用した訳語候補生成の性能（英→日）

カテゴリ名	正解訳語が n 位以内		正解生成可/候補生成可		
	$n=1$	$n=10$	○/○	×/○	×/×
電磁気学	27%	30%	30%	18%	52%
電気工学	29%	42%	42%	31%	27%
光学	32%	45%	45%	6%	48%
プログラム言語	34%	48%	55%	24%	21%
プログラミング	41%	45%	52%	24%	24%
コンピュータ	29%	34%	34%	34%	32%
解剖学	40%	50%	55%	18%	27%
疾患	21%	29%	30%	18%	52%
化学物質	20%	26%	27%	16%	57%
物理化学	31%	40%	41%	31%	28%
加重平均	29%	37%	39%	23%	38%

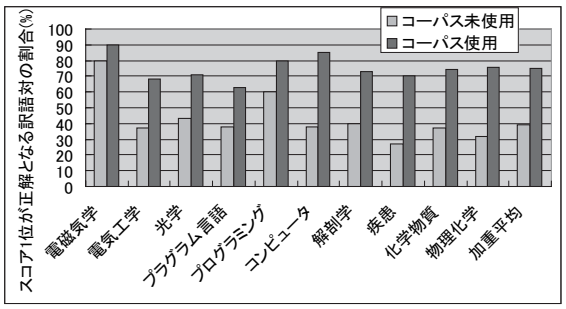

図 4 コーパスに正解が存在し、要素合成法で生成可な評価用訳語対のみを対象とした評価 (英→日)

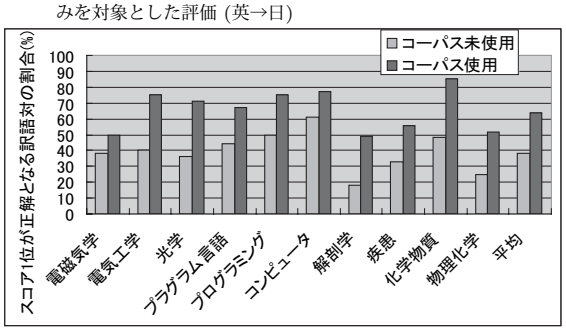

図 5 コーパスに正解が存在し、要素合成法で生成可な評価用訳語対のみを対象とした評価 (日→英)

がって、コーパスによる検証は有効であるといえる。英日方向に比べて、日英方向の精度向上が少ないのは、英語のコーパス (平均 390MB) は日本語のコーパス (平均 99MB) に比べて大きく、誤った訳語候補がコーパス中の語と照合してしまう割合が大きいためと考えられる。

次に、1 位にランクされた訳語候補の精度、すなわち、出力される訳語候補の数が 1 以上の訳語対のみを対象とした評価を行った。

$$\text{精度} = \frac{\text{コーパス (未使用/使用) で 1 位が正解}}{\text{コーパス (未使用/使用) で、候補生成可}} \quad (4)$$

その結果を図 6(英→日)、図 7(日→英) に示す。その結果、コーパス利用時は、英日方向で平均 47%、日英方向で平均 38%の精度であることが分かった。出力される訳語候補の数が 0 の場合には、他の方法で訳語推定をすれば、さらなる訳語推定精度の向上が期待できる。

5. まとめと今後の課題

本稿では、特定分野・トピックに対する対訳用語集を自動生成するという枠組みのもとに、要素合成法とコーパスを併用した訳語推定法を提案した。実験の結果、要素合成法によって、正解訳語がスコア 10 位以内となる割合は約 40%であり、コーパスを併用すると、スコア 1 位の訳語候補の正解率が大幅に向上した。

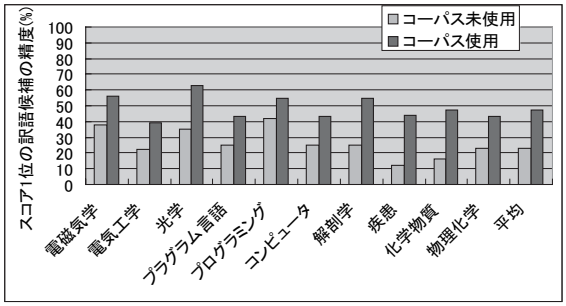

図 6 出力される訳語候補の数が 1 以上の訳語対のみを対象とした評価 (英→日)

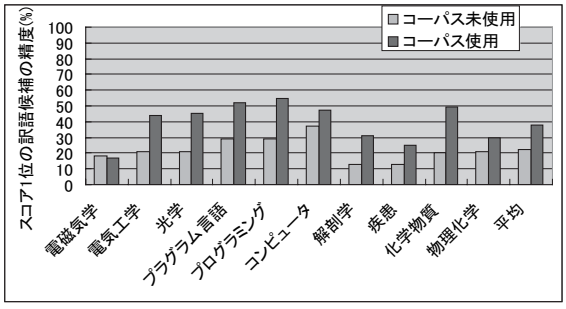

図 7 出力される訳語候補の数が 1 以上の訳語対のみを対象とした評価 (日→英)

今後の課題としては、of を含む場合などの操作の拡充が挙げられる。その他の課題としては、部分対応対訳辞書の拡充、辞書の訳語対に与えるスコアの改良などが挙げられる。今回は、コーパスによって訳語候補の検証を行ったが、サーチエンジンのヒット数を利用した検証についても検討している。

謝辞: 本研究の一部は，次の研究費による: 文部科学省 科学研究費 特定領域研究「実世界の関連性を投影した語彙空間の構築」(課題番号 16016249)，文部科学省 科学研究費 若手研究 (B) 「実世界の大規模言語資源からの翻訳知識獲得に基づく機械翻訳モデルの研究」(課題番号 16700141)，21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」.

参 考 文 献

1) Matsumoto, Y. and Utsuro, T.: Lexical Knowledge Acquisition, Handbook of Natural Language Processing (Dale, R., Moisl, H. and Somers, H.(eds.)), Marcel Dekker Inc., chapter 24, pp. 563–610 (2000).
2) 高木俊宏, 木田充洋, 外池昌嗣, 佐々木靖弘, 日野浩平, 宇津呂武仁, 佐藤理史: ウェブを利用した専門用語対訳集自動生成のための訳語候補収集, 言語処理学会第 11 回年次大会論文集 (2005).
3) 英辞郎: `http://www.eijiro.jp/`.
4) 黒橋禎夫, 長尾真: 日本語形態素解析システム JUMAN version 3.62 使用説明書 (1999).