対訳辞書とウェブを利用した専門文書中の用語の訳語推定

馬場　康夫[†]　　外池　昌嗣[††]
宇津呂　武仁[††]　　佐藤　理史[†††]

[†] 京都大学 工学部 電気電子工学科　[††] 京都大学 情報学研究科　[†††] 名古屋大学大学院 工学研究科

1. はじめに

世の中には、さまざまな分野について記された、膨大な量の多言語文書が存在する。近年のインターネットの技術の発展に伴い、私たちはそのような文書を簡単に得ることができるようになった。しかし、ある特定の分野について述べられている専門文書は、最新の技術や狭い領域の事象を対象にしているために、既存の汎用対訳辞書に存在していないような専門用語を含んでいる。このような専門用語の訳語を逐一調べることは、多大なコストを要する。

本稿では、そのような専門用語のうち 2 語以上のものに焦点を当て、ある専門文書が与えられたとき、その文書中に存在する、汎用対訳辞書に訳が載っていないような 2 語以上からなる専門用語に対して、その訳語を推定する手法について論じる。

その手法の概要を述べる。まず、専門文書中の用語のうち、汎用対訳辞書によって訳語が唯一に定まる用語を抽出する。そしてその用語の訳語を含む文書をウェブから収集し、これを専門分野コーパスとする。次に、そのコーパスを利用した要素合成法に基づき、訳語が未知な用語の訳語を推定する。

コーパスを利用した要素合成法による訳語推定に関する研究は、外池ら [1),2)] によって既に行われている。外池ら [1),2)] は、以下の方法によって要素合成法の評価を行った。まず、専門辞書のある同一分野に分類される用語を、汎用対訳辞書に訳語が載っている用語集合と、訳語が載っていない用語集合の二つに分類する。次に一定の条件に従って、それぞれの集合から、コーパス収集用の用語集合と、評価用の用語集合を選定する。最後に、コーパス収集用の用語の訳語を用いてコーパスを収集し、そのコーパスを利用した要素合成法により、評価用の用語に対して訳語推定を行う。以上のように、外池ら [1),2)] は、コーパス収集技術および要素合成法の性能評価において人工的に選んだ用語集合を用いており、これらの手法を要素技術として評価しているに過ぎない。それに対して本稿では、ある一つの文書および汎用対訳辞書が入力として与えられ、この文書中の用語のうち、汎用対訳辞書に訳語が載っていない用語の訳を推定するという、より実際的なタスクを設定する。そして、外池ら [1),2)] における手順と同様に、入力文書中の用語のうち、汎用対訳辞書に訳語が載っている用語の訳語を用いてコーパスを収集し、そのコーパスを利用した要素合成法により、汎用対訳辞書に訳語が載っていない用語の訳を推定する。評価実験の結果では、外池ら [1),2)] によって示されているものと同等以上の性能を達成しており、本稿で設定したより実際的なタスクにおいて、コーパスを利用した要素合成法による訳語推定法が適用可能であることが分かった。

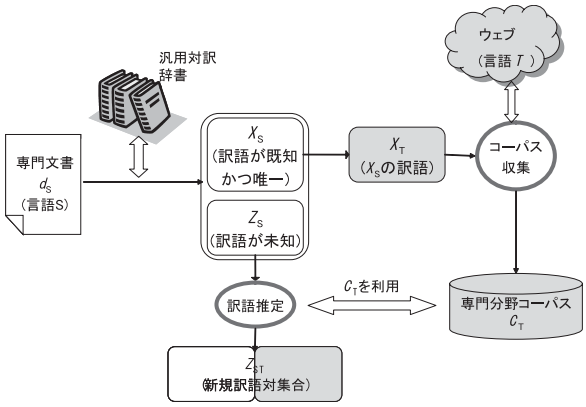

図 1　対訳辞書とウェブを利用した専門文書中の用語の訳語推定

2. 概　　要

この節では、ウェブから収集した専門分野コーパスを利用した要素合成法に基づく訳語推定について述べる。概念図を図 1 に、手順を以下に示す。

(1) 対象としている専門文書 d_S から、次の二つの用語の集合を抽出する。
- X_S：汎用対訳辞書により訳語が既知、かつ、唯一に定まる用語の集合
- Z_S：汎用対訳辞書に訳語がなく、かつ、2 語以上から構成される用語の集合

(2) 汎用対訳辞書により集合 X_S の要素をそれぞれ訳し、集合 X_T を得る。

(3) X_T の各要素からクエリを作成し、検索エンジンに与えて専門分野コーパス C_T を収集する。

(4) 収集したコーパス C_T を利用し、要素合成法により Z_S の各要素について訳語を推定する。

(5) Z_S の各要素に対応する訳語の集合 Z_T が得られる。つまり、新たな訳語対集合 Z_{ST} が得られる。

この手法の核となるのは、以下の二点である。
- 専門分野コーパスの収集
- 要素合成法に基づく訳語推定

以下の第 3 節および第 4 節では、この二点についてそれぞれ詳述する。

3. 専門分野コーパスの収集

この節では、言語 T の専門分野コーパス C_T をウェブ

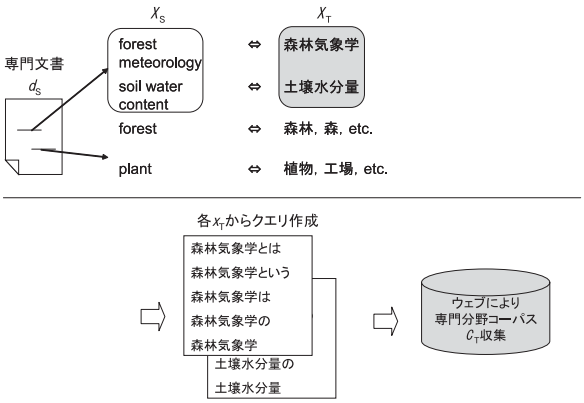

図 2 専門分野コーパス収集の概念図

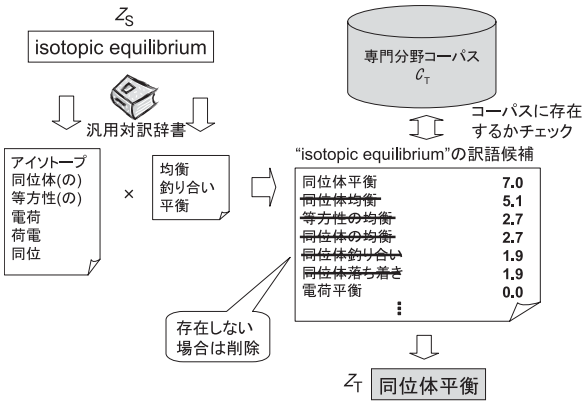

図 3 要素合成法の概念図

から収集する方法について述べる。概念図を図 2 に示す。ここで、専門分野コーパスとは、入力として与えた専門文書と内容的に深い関連性を持つような文書群をウェブから大量に収集し、それを結合したもののことである。

まず、専門分野コーパスを収集するために用いるシードの選定方法について述べる。外池ら [1),2)] はシードとして、あらかじめ与えられた用語集合を用いたが、本稿ではそのような用語集合は与えられないので、文書からシードを選定しなければならない。ここで、一般語としての性質が強い用語ほど辞書に載っている語義の数は多く、専門語としての性質が強い用語ほど語義の数は少ないという傾向がある。そこで本稿では、汎用対訳辞書に訳語がただ一つのみ載っている用語 x_S の集合 X_S を抽出し、その各要素を訳した語の集合 X_T をシードとする。

次に、X_T を用いて専門分野コーパスを収集する方法について述べる。用語 x_T ($\in X_T$) に対して、「x_T」「x_T とは」「x_T という」「x_T は」「x_T の」という 5 つのクエリを生成し、検索エンジンに与え、日本語のページから検索する。そして得られた URL の上位 100 ページ（100 ページに満たない場合は最大）をダウンロードし、さらにそのページ中で、用語 x_T がアンカーテキストになっているアンカーが存在する場合は、そのアンカー先のページも入手する [3)]。

このクエリ生成をすべての用語 x_T に対して行い、得られたページすべてを結合したものを、言語 T の専門分野コーパス C_T とする。

4. 要素合成法に基づく訳語推定

この節では、訳語が未知な専門用語 z_S の訳語 z_T を求めるために用いる要素合成法について述べる。概念図を図 3 に示す。

以下、“isotopic equilibrium” の訳語推定を例にとって説明する。汎用対訳辞書で “isotopic” という語を引くと、「アイソトープ」「同位体（の）」... という訳語が、また “equilibrium” という語を引くと、「均衡」「平衡」... という訳語が得られる。この用語が構成的であるとすれば、構成要素の訳語を組み合わせることで、“isotopic equilibrium” の訳語候補を得られるはずである。⟨ isotopic, アイソトープ ⟩, ⟨ isotopic, 同位体 ⟩ などの構成要素の訳語対には、汎用対訳辞書およびそれを加工した辞書を用いてスコアが与えられており、訳語候補のスコアは、構成要素の訳語対のスコアの積により求められる。あらかじめ出力する訳語候補数 N を設定しておけば、動的計画法により、生成過程中において N 位より下位の候補を捨てながら訳語候補を出力する。

この基本的な考え方に加え、要素合成法では、専門分野コーパスを利用した訳語候補の削除を行う。すなわち、生成過程中に現れた訳語候補が専門分野コーパス C_T に存在していない場合、その候補を削除する。

5. 実験および評価

5.1 評 価 方 法

評価実験では、対象とする文書は英語で書かれた 4 本の学術論文とする。ただし、訳語推定すべき用語に、非構成的な用語が極端に多く含まれていないという条件で選んだ。それぞれ、「農学」「地理学」「材料科学」「医学」の分野に関する内容の論文である。論文の詳細を表 1 に示す。また、求める訳語は日本語とする。訳語推定の対象とする専門用語は、「汎用対訳辞書には載っていないが専門辞書には載っている、2 語以上からなる用語」と定める。ただし、人が見て明らかに専門用語として不適格な語は除く。訳語推定の正解は、専門辞書に載っている訳語、または、人手調査の結果十分正解と見なすことのできる訳語とする。正解が複数ある場合は、そのいずれかを推定できれば良いものとする。汎用対訳辞書として英辞郎を、専門辞書としてマグローヒル科学技術用語大事典、学術用語集、医学用語大辞典、コンピュータ用語辞典の四種を用いた。辞書の詳細は表 2 に示す。

上記の方法により抽出された、訳語推定すべき専門用語の集合 Z_S の要素数を、表 3 に示す。各文書のサイズが異なるため、Z_S の大きさはまちまちである。

専門用語コーパス収集のための検索エンジンには、goo☆を用いた。

☆ http://www.goo.ne.jp

表 1 訳語推定の実験に用いる文書の詳細

分野名	詳細	
農学	題名	Seasonal dynamics and partitioning of isotopic CO_2 exchange in a C_3/C_4 managed ecosystem
	出典	"Agricultural and Forest Meteorology" Vol. 132, Issues 1-2, Pages 1-170 (20 September 2005)
	サイズ	75KB、11394 語
地理学	題名	Gender and agrobiodiversity: a case study from Bangladesh
	出典	"The Geographical Journal" Vol. 171, Issue 3, Page 195-285 (September 2005)
	サイズ	61KB、9639 語
材料科学	題名	Ferromagnetism of ZnO and GaN
	出典	"Journal of Materials Science: Materials in Electronics" Vol. 16, Number 9 (September 2005)
	サイズ	199KB、30537 語
医学	題名	The obesity hypoventilation syndrome
	出典	"The American Journal of Medicine" Vol. 118, Issue 9, Pages 935-1060 (September 2005)
	サイズ	35KB、4587 語

表 2 使用した辞書

辞書名	収録語数	出版社
英辞郎 Ver.79	129 万項目	アルク
マグローヒル科学技術用語 大辞典 第3版	11 万 6000 語	マグローヒル科学技術用語 大辞典編集委員会
学術用語集	19 万語	国立情報学研究所
25 万語医学用語大辞典	25 万語	日外アソシエーツ
コンピュータ用語辞典 第 3 版	3 万語	日外アソシエーツ

表 3 訳語推定すべき専門用語の数 $|Z_S|$

	農学	地理学	材料科学	医学	計
$\|Z_S\|$	30 個	12 個	57 個	18 個	117 個

表 4 シード X_T による専門分野コーパスの収集結果

	農学	地理学	材料科学	医学
シードの数 $\|X_T\|$	188 個	132 個	332 個	102 個
コーパス C_T のサイズ	190MB	206MB	268MB	85MB

5.2 評価結果

5.2.1 専門分野コーパスの収集の結果

まず専門分野コーパスを集めるために必要な語 X_S を抽出し、その訳語 X_T を用いて専門分野コーパス C_T を収集した。その結果を表 4 に示す。

先述したように、要素合成法では専門分野コーパスに存在している訳語のみしか訳語候補として出力することができない。したがって、収集した専門分野コーパスにおける評価用訳語 z_T の収集率が、本稿で用いる手法における要素合成法の性能の上限となる。その割合は、図

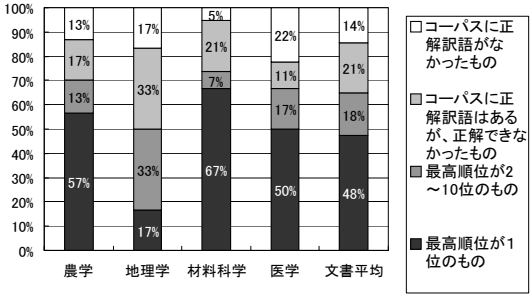

図 4 要素合成法による訳語推定の性能

表 5 訳語推定が成功した訳語対の例

	z_S	推定結果 z_T	正解順位
農学	carbon isotope ratio	炭素同位体比	1 位
地理学	local variety	地方品種	1 位
材料科学	localized magnetic moment	局在磁気モーメント	1 位
医学	upper airway obstruction	上気道閉塞	1 位

4 の「コーパスに正解訳語がなかったもの」を除く部分の和である。4 文書について収集率の平均を取ると、86%となった。これは外池ら [1] における収集率の値、78%を上回る。これが要素合成法を用いた訳語推定を行うときの正解率の上限となる。

5.2.2 要素合成法による訳語推定の結果

次に、専門分野コーパス C_T を利用した要素合成法により、各 z_S の訳語推定を行った。各文書に対する推定の性能を図 4 に示す。また表 5 に、各文書について訳語推定が成功した例を一つずつ示す。得られた訳語候補の第 1 位が正解である割合は平均 48%、10 位以内に正解が含まれる割合は平均 66%であった。

5.2.3 誤り分析

この節では、要素合成法による訳語推定に失敗した用語の誤り分析を行う。本稿で用いている要素合成法は、専門分野コーパスに存在しない用語を訳語候補として選ばないシステムなので、コーパスに正解が存在しないものは、そもそも正解できない。よって、ここでは、コーパスに正解は存在するが、訳語推定に失敗したものに焦点をあてて誤り分析を行った。その結果を表 6 に示す。

これより、訳語推定に失敗する原因の約 7 割は、「辞書に構成要素の訳語がない」「訳語対が非構成的」の二つが占めていることが分かるが、これらは要素合成法を用いる限り根本的に対処不可能である。また、その他の誤り原因は多岐に渡っているため、網羅的に対処するのは容易ではない。

5.2.4 併記された訳を利用する方法との比較

本稿では訳語を推定するのに、用語を構成要素に分解し、それぞれの訳を組み合わせるという手法を用いたが、他にも、訳語候補となりうる語をウェブから検索し、それにスコア付けを行う手法も考えられる [4),5)]。ここでは、

表 6 コーパスに正解訳語は存在するが、訳語推定に失敗したものについての誤り分析

	誤り例		個数
	z_S	z_S の正しい訳語	
辞書に構成要素の訳語がない	extensive cultivation	粗放栽培	12
訳語対が非構成的	glycine max	ダイズ	5
11 位以下			2
訳語に「の」「性」などが必要	leaf senescence	葉の老化	2
異表記	agricultural labour	農業労働	1
不明			1
合計			23

図 5 2.「スニペット」に分類される用語の例

この異なる手法がそれぞれカバーできる用語の範囲の違いについて調べるために、併記された訳を利用する方法による以下のような予備的な実験を行った。

(1) goo のオプションで言語を日本語のみに設定し、z_S を検索して、上位 20 個のスニペット☆を得る。

(2) そのスニペットから各 z_S の訳語候補がどの程度得られるかを人手により調査し、その難易度によって以下の 4 つに分類する。

1. 併記：z_S の近くに括弧付けで訳語が併記されており、自信を持って訳語であることが分かる場合
2. スニペット：分類 1. 以外で、スニペットに訳語が存在していて、人が見て推論可能な場合（例として “genetic resources” を検索して得られたスニペットを図 5 に示す。これにより訳語「遺伝資源」が推論可能である。）
3. リンク：スニペットには訳語がないが、リンク先を一段階辿ったページを見れば分類 2. と同じように推論可能なもの
4. 失敗：訳語が見つからないもの

この手順により手作業で得た訳語候補を分類した結果を図 6 に示す。1.〜3. に分類された用語に関しては、訳語を自動推定できる可能性がある。

次に、併記された訳を利用する方法で 4. に分類された z_S について、要素合成法での正解率を調べた。その結果を表 7 に示す。この結果から、併記された訳を利用する方法で訳語候補を得られなかった用語の中には、要素合成法を用いれば訳語を推定可能なものが一定の割合で存在することが分かった。このことは、訳語推定法と訳語併記を利用する方法がまったく異なる手法のために、それらを組み合わせることでそれぞれの欠点を補完し合える可能性があることを示唆している。

☆ 検索時に得られる短い説明文のこと。

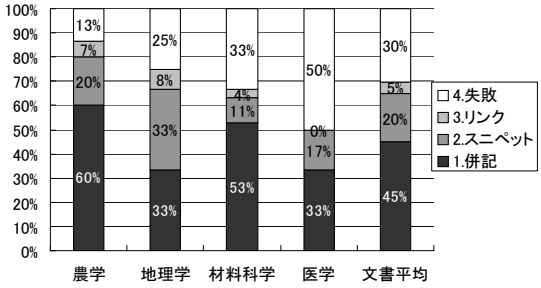

図 6 併記された訳を利用する方法により手作業で得た訳語候補の分類

表 7 訳語併記を利用する方法で訳語を発見できなかった用語に対する、要素合成法での正解率

	併記の方法で 4. に分類された用語の個数 (*)	(*) のうち、要素合成法では 1 位		(*) のうち、要素合成法では 10 位以内	
農学	4	4	100%	4	100%
地理学	3	0	0%	1	33%
材料科学	19	13	68%	14	74%
医学	9	3	33%	5	55%

6. ま と め

本稿では、ある専門文書が与えられたとき、その専門文書中に存在する、汎用対訳辞書に訳が載っていないような 2 語以上からなる用語に対して、収集した専門分野コーパスを利用した要素合成法によって訳語を推定する手法について論じた。得られた訳語候補の第 1 位が正解である割合は平均 48%、10 位以内に正解が含まれる割合は平均 66%であった。

謝辞: 本研究の一部は，次の研究費による：国立情報学研究所共同研究「単言語・多言語環境で再利用可能な言語単位の高度活用手法に関する研究」.

参 考 文 献

1) 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).
2) 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).
3) 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).
4) Nagata, M., Saito, T. and Suzuki, K.: Using the Web as a Bilingual Dictionary, *Proc. Workshop on Data-driven Methods in Machine Translation*, pp. 95–102 (2001).
5) Huang, F., Zhang, Y. and Vogel, S.: Mining Key Phrase Translations from Web Corpora, *Proc. HLT/EMNLP*, pp. 483–490 (2005).