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*

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[††]
[†]京都大学 工学部 電気電子工学科　[††]京都大学 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

一般に，技術翻訳や同時通訳などの翻訳の分野においては，多様な専門的分野にわたって専門用語に関する翻訳知識が用いられる．そのような多様な分野の専門用語の訳語の情報は，汎用辞書 (例えば，英辞郎 http://www.alc.co.jp) に含まれていないものが多い．しかし，多様な分野について，人手で専門用語対訳集を作成するためには多大なコストを必要とする．そこで本研究では，特定分野 · トピックの対訳用語集を自動生成することを目的とする．

対訳コーパスやコンパラブルコーパスの中に含まれる用語の訳語を獲得する手法の研究は，従来より行われてきた ([Matsumoto00])．ところがこれらの手法では，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．これらの従来の訳語獲得の研究と比べると，特定分野 · トピックの対訳用語集自動生成においては，(1) 特定分野 · トピックのコーパスがあらかじめ与えられているとは限らず，いかにしてこれを収集するか，(2) 収集したコーパスから，その分野 · トピックに関する専門用語のリストをいかにして作成するか，という課題を解決する必要がある．

そこで，本稿では，特定分野 · トピックの対訳用語集自動生成の問題に対して，言語 S において特定分野 · トピックの用語サンプルが与えられたとして，言語 S, T の間で，専門用語対訳集を自動生成するという形式のタスクを設定する．この流れを図 1 に示す．まず与えられた用語サンプルに対して [佐々木 04] の関連語収集手法を適用することにより，その分野 · トピックの用語（言語 S）を収集する．収集したそれぞれの用語を汎用対訳辞書と照合し，辞書に訳が載っている用語 x_S の集合 X_S と，訳が載っていない用語 y_S の集合 Y_S に分ける．そして，訳が載っていない用語 y_S に対して，その訳語 y_T を推定する．訳語 y_T を推定するにあたっては，まずその候補を収集する．ここでは，ウェブ上の検索エンジンを利用して，用語 x_S の訳語 x_T と関連の強いページを収集し，これを特定分野 · トピックに関連した言語 T のコーパスとみなす．そして，このコーパスから訳語 y_T の候補を抽出し，この訳語候補と用語 y_S との間で訳語推定を行う [外池 05]．

このような枠組みのもとで，特に，本稿では以下の手順により訳語候補収集を行うこととし，その性能を評価する．

> 用語集合 X_S，それらの用語に対する訳語の集合 X_T，および，訳語が未知である用語集合 Y_S が与えられたとして，検索エンジンを利用して，言語 T におけるコーパスを収集し，そのコーパスから，用語集合 Y_S 中の用語 y_S の訳語候補を収集する． (1)

2　特定分野・トピックのコーパス収集

図 1 の特定分野 · トピックの対訳用語集自動生成の流れにおいて，言語 T のコーパスを収集する手順を以下に述べる．まず，用語 x_T を用いてクエリを生成し，これを検索エンジンに入力して得られた URL のそれぞれ上位 100 ページをコーパスとする．それらのページに，用語 x_T がアンカーテキストとなっているアンカーが存在する場合は，そのアンカー先ページも入手する [佐々木 04]．

言語 T が日本語の場合，用語 x_T に対して「x_T とは」「x_T という」「x_T は」「x_T の」「x_T」という 5 種類のクエリを作成し，英語の場合は,「x_T」「x_T　AND　what's」「x_T　AND　glossary」という 3 種類のクエリを作成する．また，検索エンジンは，日本語は goo[1] を，英語は AltaVista[2] をそれぞれ用いる[3]．そしてそれぞれの x_T に対して収集したコーパスすべてを結合したものを，その分野 · トピックにおける言語 T のコーパスとする．

*Collecting Translation Candidates from Web in Automatic Generation of Bilingual Glossary for Technical Terms

[1] http://www.goo.ne.jp/
[2] http://www.altavista.com/
[3] 以下，用語のヒット数を調べるときも同様である．

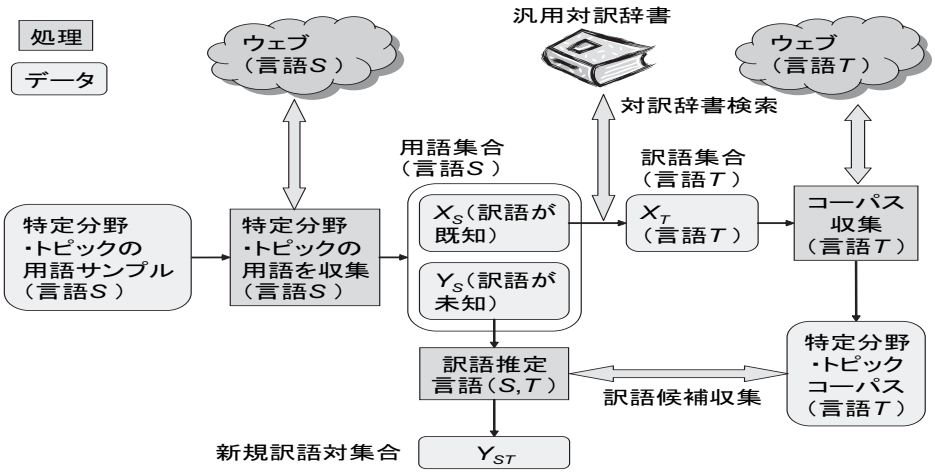


図 1: 特定分野 · トピックの対訳用語集自動生成の流れ

3 実験および評価

3.1 評価方法

本稿の評価実験においては，以下の手順により式 (1) の過程を評価する．汎用対訳辞書に含まれる訳語対の集合を既知訳語対集合 X_{ST}，含まれない訳語対の集合を評価用訳語対集合 Y_{ST} とする．また，X_{ST} に含まれる言語 T の用語 x_T の集合を X_T，Y_{ST} に含まれる言語 T の用語 y_T の集合を Y_T とする．既存の専門用語対訳辞書から X_{ST} および Y_{ST} を選定し，X_T の要素 x_T に対して 2 節で述べた方法を適用することにより，言語 T のコーパスを収集する．そして，X_T の全要素を用いて収集したコーパスに Y_T の要素 y_T が含まれる割合を調べる．

3.2 評価用訳語対

本稿の評価実験では，言語 $\langle S, T\rangle$ の組を 〈 日本語, 英語 〉 または 〈 英語, 日本語 〉 とし，4 種類の日英対訳用語辞書から選んだ 10 分野 (表 1) の日英訳語対を対象として，既知訳語対集合 X_{ST} および評価用訳語対集合 Y_{ST} を選定した．

X_{ST} および Y_{ST} の選定は，具体的には次のようにして行った．日英対訳用語辞書の訳語対を $\langle z_S, z_T\rangle$ とする．まず，z_S および z_T それぞれを検索エンジンで検索したときのヒット数を求める．z_S, z_T がともにヒット数 100 以上であった訳語対 $\langle z_S, z_T\rangle$ に対して，これを汎用訳語辞書 (英辞郎 ver.79（129 万語）) と照合し，$\langle z_S, z_T\rangle$ が汎用対訳辞書に含まれているものを既知訳語対，$\langle z_S, \cdot\rangle$，$\langle \cdot, z_T\rangle$ いずれも含まれていないものを評価用訳語対とした．なお，ヒット数に関する訳語対の選定基準として，既知訳語対はその英語側の語のヒット数が 1,000 以上 10,000 以下のもの，評価用訳語対は英語側の語のヒット数が 10,000 以下のものとした．表 1 の分野について選定した X_{ST} および Y_{ST} それぞれの訳語対数を表 2 に示す．

ここで注意すべき点として，選定した分野の大きさは一様ではないことが挙げられる．例えば，分野「コンピュータ」はコンピュータ分野全般に対応しているのに対して，分野「プログラム言語」や「プログラミング」はその中の小分野と考えられる．このような分野の大きさの違いは，以下の実験の結果に影響すると考えられる．

表 1: 評価実験に用いた対訳辞書および分野

英日対訳用語辞書	収録語数	分野
マグローヒル 科学用語大辞典	11 万 6 千語	電磁気学 電気工学 光学
岩波情報科学辞典	用語の木から 選定	プログラム言語 プログラミング
コンピュータ用語辞書	3 万語	コンピュータ
25 万語医学用語大辞典	25 万語	解剖学 疾患 化学物質及び薬物 物理化学及び統計学

表 2: 評価に用いた訳語対の数

分野	既知訳語対数	評価用訳語対数
電磁気学	73	33
電気工学	72	45
光学	71	31
プログラム言語	61	29
プログラミング	63	29
コンピュータ	100	100
解剖学	100	100
疾患	100	100
化学物質及び薬物	100	100
物理化学及び統計学	100	100

3.3 特定分野のコーパス

既知訳語対集合 X_{ST} の言語 T の用語 x_T を用いて収集した各分野のコーパスのサイズを表 3 に示す (項目

表 3: ヒット数上限値により既知訳語対数を減少させたときの評価用訳語の収集率

(a) 日本語

分野	ヒット数上限	既知訳語対数の比 (%)	コーパスサイズ (Mbytes) の比 (%)	評価用訳語の収集率の比 (%)
電磁気学	1000	42 (31/73)	31.6 (31.3/99.0)	97 (82/85)
電気工学	4000	74 (53/72)	62.4 (48.2/77.3)	92 (67/73)
光学	2000	61 (43/71)	43.3 (39.8/92.0)	92 (71/77)
プログラム言語	2000	57 (35/61)	46.8 (38.7/82.8)	92 (79/86)
プログラミング	3000	46 (29/63)	38.4 (40.5/106)	93 (86/93)
コンピュータ	3000	86 (86/100)	79.9 (86.3/108)	96 (47/49)
解剖学	1000	70 (70/100)	63.1 (73.4/116)	95 (74/78)
疾患	1000	65 (65/100)	52.4 (65.1/124)	95 (76/80)
化学物質及び薬物	1000	73 (73/100)	63.7 (59.7/93.7)	96 (71/74)
物理化学及び統計学	2000	89 (89/100)	85.1 (79.3/93.2)	92 (56/61)

(b) 英語

分野	ヒット数上限	既知訳語対数の比 (%)	コーパスサイズ (Mbytes) の比 (%)	評価用訳語の収集率の比 (%)
電磁気学	50000	55 (40/73)	68.6 (193/281)	97 (85/88)
電気工学	50000	57 (41/72)	64.0 (153/240)	100 (76/76)
光学	8000	32 (23/71)	29.6 (75.8/256)	91 (74/81)
プログラム言語	8000	9.8 (6/61)	10.0 (39.2/392)	90 (90/100)
プログラミング	30000	29 (18/63)	32.2 (108/337)	96 (93/97)
コンピュータ	8000	81 (81/100)	79.7 (506/634)	94 (63/67)
解剖学	3000	68 (68/100)	63.0 (263/418)	92 (66/72)
疾患	5000	75 (75/100)	69.2 (301/435)	94 (64/68)
化学物質及び薬物	7000	73 (73/100)	63.7 (286/448)	90 (56/62)
物理化学及び統計学	8000	81 (81/100)	77.2 (351/454)	93 (62/67)

「コーパスサイズ (Mbytes) の比 (%)」の括弧内の分母の値). また，それぞれの分野において，収集した URL 数は日英ともおおよそ 10,000 ページ前後であり，そのうち異なりページの割合は，英語で 88~97%, 日本語で 92~99%であった.

3.4 評価用訳語の収集率

評価実験においては，既知訳語対集合 X_{ST} に含まれる用語 x_T を用いて収集したコーパスから得た訳語候補集合の中に，評価用訳語対集合 Y_{ST} に含まれる用語 y_T が出現する割合（具体的には，以下の収集率）を調べることによって，その収集性能を評価した.

$$\text{評価用訳語の収集率} = \frac{|\{y_T \in Y_T | \, y_T\text{はコーパス内に出現}\}|}{|Y_T|} \quad (2)$$

ここでは，特に，用語 y_T の出現頻度下限 L_f を 2 および 5 として収集率を評価した．言語 T が日本語の場合，および英語の場合の結果をそれぞれ図 2(a),(b) に示す．日本語，英語どちらの場合も岩波情報科学辞典の分野の収集率が高く，一方，コンピュータの収集率は低くなった．これは 3.2 節でも述べたように，分野によってその大きさが異なっており，訳語対の分布の大きさが異なることが原因と考えられる.

次に，集めたコーパスが，その分野に関連する情報を含んだ良質な内容を集められているかを調査するため，ある分野の用語集合 X_T から集めたコーパスに対する，異なる分野の用語集合 Y_T の収集率を調べた．その結果を表 4 に示す.

「コンピュータ」のコーパスに対して他の分野の訳語候補収集率が高いのは,「コンピュータ」の分野が大きいことを示していると考えられる．(異分野の収集率)>(同一分野の収集率) となった組合せは，お互い類似した分野同士であるものが多い．一方,「プログラム言語」のコーパスに対する医学系の訳語候補の収集率が低いことを考えると，(1) 関連が低い分野同士では収集率が低いことから，分野に特化した内容のコーパスが作成できている，(2) 意味の狭い分野においては，そのコーパスも狭い範囲に絞られている，と言える.

3.5 既知訳語対数と訳語候補収集率の相関

ここまでの実験における問題点として，言語 T において作成したコーパスのサイズが非常に大きく (3.3 節)，計算時間を大幅に消費する点が挙げられる．そこで，コーパス作成に用いる既知訳語対数を変化させて訳語候補収集率との相関を調べた.

既知訳語対集合 X_{ST} に含まれる用語 x_T のそれぞれに対して検索エンジンを用いてその用語のヒット数を求める．そして，上限値以下のヒット数を持つ用語 x_T のみを用いて作成したコーパスにおける訳語候補収集率を求めた．全ての x_T を用いたときの収集率に対して，その 90%を下回らない収集率となるときのヒット数上限値，コーパスサイズ，既知訳語対数，および評価用訳語の収集率を表 3 に示す．表 3 から分かるよう

表 4: 評価用訳語の収集率 (%)(同一分野、異分野)

(a) 日本語

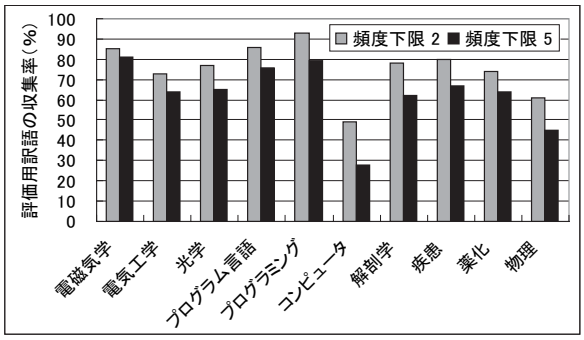
		分野 (評価用訳語)									
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
分野 (コーパス)	(1) 電磁気学	**85**	71	48	17	24	28	10	4	13	49
	(2) 電気工学	76	**73**	29	17	21	16	6	5	6	43
	(3) 光学	45	13	**77**	10	17	11	13	13	8	28
	(4) プログラム言語	27	29	6	**86**	83	34	2	2	1	9
	(5) プログラミング	18	24	10	93	**93**	37	5	4	5	9
	(6) コンピュータ	58	56	29	69	76	**49**	4	4	6	35
	(7) 解剖学	0	4	19	14	21	5	**78**	67	30	31
	(8) 疾患	3	2	3	3	0	2	65	**80**	30	20
	(9) 化学物質及び薬物	6	13	6	0	0	4	30	39	**74**	31
	(10) 物理化学及び統計学	55	51	42	34	34	19	22	11	38	**61**

下線部：異分野の収集率 ＞ 同一分野の収集率

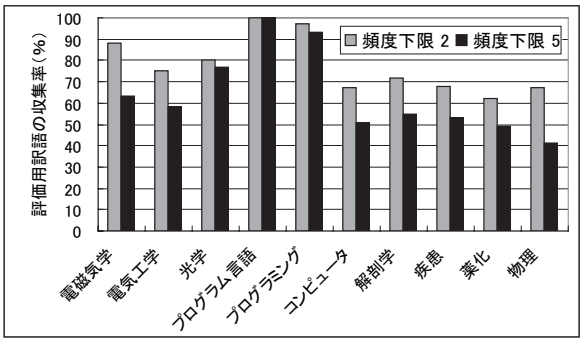
(b) 英語

		分野 (評価用訳語)									
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
分野 (コーパス)	(1) 電磁気学	**88**	62	74	62	31	26	11	4	10	41
	(2) 電気工学	67	**76**	48	45	34	25	8	6	8	42
	(3) 光学	52	42	**81**	41	28	20	13	7	16	38
	(4) プログラム言語	12	20	6	**100**	93	46	5	0	8	20
	(5) プログラミング	18	22	13	100	**97**	48	7	3	6	20
	(6) コンピュータ	45	42	39	100	83	**67**	9	5	11	31
	(7) 解剖学	9	18	16	38	24	6	**72**	56	33	34
	(8) 疾患	6	9	6	21	10	8	54	**68**	35	21
	(9) 化学物質及び薬物	12	20	23	21	24	11	31	31	**62**	40
	(10) 物理化学及び統計学	61	58	74	55	55	31	23	24	45	**67**

下線部：異分野の収集率 ＞ 同一分野の収集率



(a) 日本語



(b) 英語

図 2: 評価用訳語の収集率

に，訳語候補を十分に収集できるコーパスを集めるために必要な用語 x_T は，実質的にはそのヒット数が大きくないもので十分であるといえる．表 3 における縮小したコーパスで用いた URL の異なりページの割合は，英語で 90∼99%, 日本語で 95∼99%であり，すべての用語 x_T を用いて作成したコーパスに比べて少し増えている．また，異分野収集率は，縮小したコーパスにおいてもその傾向に変化はなかった．

4 おわりに

本稿では，特定分野・トピックの対訳用語集を自動生成する枠組みの中で，特に，言語 S の用語集合 X_S，それらの用語に対する言語 T における訳語の集合 X_T，および，訳語が未知である用語集合 Y_S が与えられたときに，検索エンジンを利用して言語 T におけるコーパスを収集し，そのコーパスから用語集合 Y_S 中の用語 y_S の訳語候補を収集する手法を提案し，評価実験によってその性能を評価した．

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

[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).