QRpotato: 専門用語対訳対の網羅的な収集

阿辺川 武　　影浦 峡
東京大学大学院 教育学研究科
{abekawa,kyo}@p.u-tokyo.ac.jp

1　はじめに

翻訳においてある専門分野の文書を翻訳するときは，その分野で普段使用されている専門用語を使用する必要がある．しかし翻訳者は，すべての専門分野に精通しているわけではないため，専門外の分野を翻訳することになった場合，その分野の専門用語集を参照しなければならない．翻訳者がその都度，翻訳する専門分野の用語集を用意するのは非常にコストがかかるという問題がある．

我々は翻訳支援システム QRedit[1] を開発しており，本システムでは原文の各用語について自動的に辞書引きを行ない，翻訳者に提示する機能を有している．辞書の数を増やすことで，それだけ多くの用語の辞書引きを行なえるが，市販の辞書を誰にでも使用できるシステムに組み込むことはライセンス上不都合なことが多い．そこで本研究では，誰もが自由に使用できる専門用語辞書の構築をめざし，Web 上で公開されている専門用語集を網羅的に収集することをめざす．本稿では，その収集段階の報告を行なう．

Web 上から特定のフォーマットで出現する情報を収集する研究は，数多く行なわれている．例えば久光ら [3] は括弧表現を用いて有用な略語と正式名称といった表現対を収集している．一方で翻訳リソースの構築の一部として対訳コーパスから訳語対を統計的指標から自動抽出しようという研究もあるが，辻ら [5] は，対訳コーパスにおいて低頻度語の抽出は難しいことを述べている．また Gey らは [2]，論文抄録の一部に含まれる，論文のキーワード覧から専門用語対訳対を収集している．

本稿では，専門用語からなるシード対訳対を利用して，Web 上でその出現ページを検索し，同一ページ内でシード対訳対と同じフォーマットで出現する文字列対を対訳候補として網羅的に収集する手法を提案する．

2　専門用語対訳対の存在するページ

最初に，Web 上で専門用語対訳対がまとめて存在する可能性が高いページの例を紹介する．

単語集リスト
Web 上には特定の概念を持つ単語対の集合を編集したページが数多く存在する．その中にはある領域の専門用語をまとめたページがある一方で「中学 3 年で習得する単語リスト」といった一般用語に関してリストを作成している場合もあり，これらと区別する必要がある．

専門文書
専門文書 (technical document) では文章中で専門用語 (technical term) を使用する場面は頻繁にあるが，対応する英語表現が存在するときその用語を併記することがある．

目次，索引
目次，索引でも同様に見出し語に対し，英語が併記されていることがある．特に学術系のコンテンツに多く，大学のシラバスも同様の傾向がある．

3　収集手法

本節では，専門用語対訳対を収集する手法を説明する．以下に収集の簡単な流れを述べる．

1. シード対訳対集合を用意する．
2. シード対訳対をクエリーとして，検索エンジンから対訳対を含むページを収集する．
3. 収集されたページごとにシード対訳対の出現フォーマットを求める．
4. 得られたフォーマットを用いて，同様のフォーマットで出現する用語対を収集する．

表 1: 得られた文字列パターン

	左側終端記号	日本語用語	中間文字列	英語用語	右側終端記号
1	’、’	シェーグレン症候群	’(’	Sjogren syndrome	’)’
2	’>’	ラポール	’ ’	rapport	’と’
3	’[’	代謝	’]](’	metabolism	’)’
4	’●’	アンダーカット	’【’	undercut	’】’
5	’>’	アンタゴニスト	’’	antagonist	’<’
6	’>’	イベント	’ (<i>’	event	’<’
7	’>’	光	’</td><td><font>’	light	’<’

3.1 専門用語対訳対をクエリーとした Web 検索

Web 検索エンジンを用いて，シードとなる専門用語対訳対を含むページを検索する[1]．その際，対訳対がその順番で隣接して出現するように，そのまま文字列を結合したフレーズで検索を行なう．そして，検索結果から得られた URL から HTML ファイルをクロールする．

3.2 対訳対の抽出

3.2.1 中間文字列の取得

既存の検索エンジンの多くは，HTML タグを含め記号全般を除去した検索インデックスを構築しており，フレーズ検索を行なっても，間に存在している記号類はすべて無視した文字列として検索される．そこで，シード対訳対の間に含まれる中間文字列を取得するため，得られた HTML ファイルについて，プログラムを用いてもう一度シード対訳対による検索を行なう．ここでは，完全なフレーズとしてではなく，対訳対の間に記号文字 (空白を含む) からなる文字列の挿入を許した正規表現検索をする．これにより表 1 にあるような中間文字列が取得できる．

3.2.2 終端記号の検索

次に，左側の用語，右側の用語双方に対して，終端文字列を決定する．中間文字列内に括弧記号や HTML タグがある場合には，それらに対応する終端記号を決定する．例えば，中間文字列に’(’ があったとき，右側の用語の終端記号は’)’ となり，’]’ がある場合は，左側の用語も終端記号は’[’ となる．

中間文字列に括弧がないときは，用語から終端方向へ向けて記号文字を検索し，最初の記号文字を終端記号とする．これは中間文字列に HTML のタグがあるときなどが当てはまる (表 1 の 4,5,6,7)．終端記号方向へ記号文字を探しているとき，別の文字種が出現したとき (表 1 では 1 の左側，2 の右側) は，ここで探索を打ち切り，文字種を用いた正規表現としてパターンを作成する (1 では「’, ’ を除いた日本語文字列」，2 では「アルファベットと数字からなる文字列」)．このため文章中に用語が出現した場合など，専門用語の区切りが判断できないことがあり，このパターンの信頼度が低くなる．

3.2.3 同一フォーマットの検索

上記で得られたパターンについて，中間文字列と終端記号から正規表現を作成し，そのページ内で同一のフォーマットで出現する文字列対を検索する．これは 1 つのページ内では，専門用語対訳対が同一のフォーマットで出現する傾向が高いというヒューリスティックスを利用している．獲得された文字列対については，どのようなパターンを用いて検索されたかについて保存しておく．前節で述べたように端が終端記号で抑えられるものと，正規表現の境界であるものを区別するためである．

4 収集したデータ

学術用語集[2] から 23 専門分野，210,328 対をシード対訳対として用いた．収集したデータの集計を表 2 に載せる．そして表 3 に獲得した対訳対候補の頻度上位 10 対を掲載する．表 4 は，シード対訳対を多く含むページのタイトルと新規獲得対訳対の数を示したものである．表 1 は，横軸にページ内に含まれる対訳対候補の数，縦軸にその候補数を持つページの数をプロットしたものである．両軸とも対数をとるとほぼ直線上に分布し，Zipf’s の法則に従っていることがわかる．

[1]検索エンジンには Yahoo!検索 Web サービス http://developer.yahoo.co.jp/webapi/search/ を使用した．

[2]国立情報学研究所編纂学術用語集 http://sciterm.nii.ac.jp/cgi-bin/reference.cgi

表 4: シード対訳対を多く含むページ

ページタイトル	シード対訳対数	新規対訳獲得対数
第３版までの第１６章　「学術的教養：学部別語彙表現」	2,026	668
和英医学用語集（内科学会 1993 ＋循環器学会 1995 ＋生理学会 1987 ）	1,928	19,079
和英医学用語集（内科学会 1993 ＋循環器学会 1995 ＋生理学会 1987 ）	1,870	19,095
日本救急医学会・用語集 1(和英)	1,345	2,146
金属加工常用語中日英対照表 A	996	686
専門用語	919	1,987
MEDO 歯科用語 (2008/03/15)	884	1,886
s95936 辞書	827	307
辞書 www.tool-tool.com–BW 數控刀具科學館	735	312
unsaved:///newpage1.htm	659	354

表 2: 収集したデータの集計

ヒットした URL の総数	1,425,107
うち HTML を取得できたページ	1,327,180
1 つ以上新規対訳対候補があったページ	893,103
総獲得対訳対候補数	6,567,186
うち重複除く	3,486,125

表 3: 頻度の多い対訳対候補

日本語候補	英語候補	出現回数
ウィキペディア	Wikipedia	13067
ブログ	Blog	2160
島	Island	1948
森	Forest	1879
山	Mountain	1850
沼	Swamp	1832
平地	Plains	1787
ホーム	SNS	1637
アドレス	URL	1480
楽天ブログ	Blog	1283

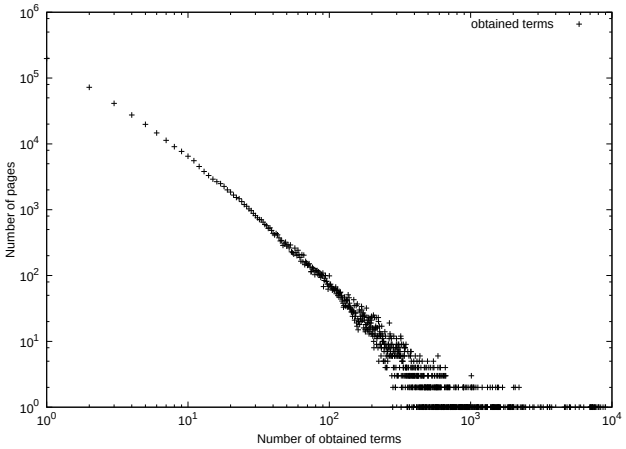

図 1: 対訳対候補数とページ数の関係

5　評価

市販の専門用語辞書[3]を利用して，収集した対訳対がどの程度専門領域を網羅しているかを調査した．調査対象とする分野はシード対訳対に存在する専門分野「農学」「医療医学」「地球環境」そして存在しない専門分野「コンピュータ」「会計」「生産工学」を選択した．表 5 がその結果である．

表 5 の結果より，シード対訳対に存在する分野は比較的が再現率の高いが，存在しない分野では低い値となった．網羅性のある専門用語対訳集をめざすためには，さらなる対訳対候補の収集が必要であり，今後，以下のことについて取り組む予定である．

今回はクエリーには「日本語 英語」の順番を用いた．今後これを逆にした「英語 日本語」，次に，読みを間に挿入した「日本語 読み 英語」，そして HTML だけでなく PDF フォーマットのファイルも収集する予定である．PDF ページで公開される対訳用語集は多く，網羅的な収集をめざすにあたって必須であると考えている．

新規に得られた対訳対候補をシード対訳対にして，同じ収集作業を繰り返す．これによりシード対訳対に存在しない専門分野に対しても多くの対訳対が収集できると期待できる．ただしゴミを含んだ状態で新たにシードを生成してもさらにゴミを増やす結果になるので，この段階の前に精度の高いフィルタリングが要求される．

本格的な誤り解析はまだ行なっていないが，専門用

[3]株式会社クロスランゲージの専門用語辞書を利用した．

表 5: 既存辞書に対する再現率 (シード対訳対を含む)

		再現率		
	語彙数	日本語見出しのみ	英語見出しのみ	対訳対
農学	9,620	0.932	0.979	0.896
医療医学	140,584	0.512	0.665	0.360
地球環境	38,950	0.400	0.385	0.240
コンピュータ*	153,659	0.287	0.321	0.176
会計*	9,091	0.201	0.097	0.031
生産工学*	2,905	0.186	0.157	0.074

*はシード対訳対には存在しない専門分野

語対訳対でない対を多く含むページには次のようなものが挙げられる．

読みだけのページ
カタカナからなる日本語用語に対応する英語用語が，たまたまローマ字に読みと同じとき (例，グアノ:guano) に，日本語の読みを列挙しているページを獲得してしまうことがある．特に日本の文化を海外に紹介するページに多い．形態素解析器を用いて読みを取得しローマ字に変換したものと，得られた英語用語を比較すれば，これらのページを除去できると思われる．

他の言語のページ
今回用いたシード用語対は日英のものであるが，たまたま英語の用語が他の言語と同じスペルであったために検索でヒットしてしまったことが原因である．対応策として，その言語に対応した一般辞書が用意できれば，新たに獲得した用語について，その言語特有の単語を含むかどうかで判断できると思われる．

6 おわりに

本稿では，Web 上から専門用語対訳対を網羅的に収集するために，シード対訳対を含むページを獲得し，ページ中でシード対訳対と同じフォーマットで出現する文字列対を新たな対訳対候補として獲得した．獲得した対訳対集合を既存の専門用語辞書との比較した結果，半分以上の再現率を示す分野がある一方，ほとんど獲得できない分野もあった．

今後は，さらなる対訳対収集へ向けた作業を継続するとともに，対訳対でない文字列対や専門用語以外の対訳対を，要素合成法 [4] などを用いた手法で信頼度を計算しフィルタリングする手法や，得られた専門用語対訳対の分野別のクラスタリングなどについて研究する予定である．

付記

本研究の一部は，日本学術振興会科学研究費補助金基盤 (A)「翻訳者を支援するオンライン多言語レファレンス・ツールの構築」(課題番号 17200018) の支援を得て行われた．

参考文献

[1] Takeshi Abekawa and Kyo Kageura. QRedit: An integrated editor system to support online volunteer translators. In *Digital humanities*, pp. 3–5, 2007.

[2] Fredric Gey, David Kirk Evans, and Noriko Kando. A japanese-english technical lexicon for translation and language research. In *LREC2008*, pp. 26–30, 2008.

[3] 久光徹, 丹羽芳樹. 統計量とルールを組み合わせて有用な括弧表現を抽出する手法. 情報処理学会研究報告 122-NL-17, pp. 113–118, 1997.

[4] 外池昌嗣, 宇津呂武仁, 佐藤理史. ウェブから収集した専門分野コーパスと要素合成法を用いた専門用語訳語推定. 自然言語処理, Vol. 14, No. 2, pp. 33–68, 2007.

[5] 辻慶太, 芳鐘冬樹, 影浦峡. 対訳コーパスにおける低頻度語の性質 : 訳語対自動抽出に向けた基礎研究. 電子情報通信学会技術研究報告, NLC2000-16, pp. 47–54, 2000.