英和／和英辞典を用いた英語換言語の抽出

鷹尾 和享　　今村 賢治　　柏岡 秀紀
ATR　音声言語コミュニケーション研究所
E-mail: {kazutaka.takao, kenji.imamura, hideki.kashioka}@atr.co.jp

1　はじめに

機械翻訳において、訳語の多様性の問題がある。当研究所の機械翻訳システムでは、初期段階で構築を急いだために対訳エントリを１対１にしたことも一因となり、１つの原言語から得られる目的言語の訳語は画一的であるという傾向がある。人間が日本語を英語に訳す場合、和英辞典の規模が小さくて適切な英訳が載っていなければ、自分の知っている英単語をいくつか思い浮かべ、逆方向に英和辞典を引き、今訳そうとしている文に最適の英語を見つけ出すという作業を行うことができる。筆者らは市販の和英辞典の電子データを機械翻訳用に利用しているが、このことを考えると、単に和英辞典のエントリを取り込むだけでは、表現の豊かさという点で不十分である。上記と同様の処理をコンピュータで行うためには、似たような意味を持つ英単語の組、すなわち、英語の換言語を収集することが必要である。

筆者らは、学研のスーパーアンカー英和辞典[1]とニューアンカー和英辞典[2]の電子データから対訳エントリを抽出した。抽出した語数は（表１）の通りである。これを見ると、見出し語数については和英辞典の方がかなり少ないのみならず、見出し語１語あたりの訳語数（B/A）については英和辞典が3.0 語であるのに対し、和英辞典は 1.6 語である。この数字から、この和英辞典の英語の訳語数は少なすぎることが確認できる。もともと人間用の和英辞典なので、利用者に対して直訳調の英語表現を避けることを促すため、意図的に訳語数を減らして用例を多くするという編集方針に起因すると思われるが、訳語をもっと豊富に得たい場合は何らかの対策を考える必要がある。

筆者らは、[3]で和英辞典と英和辞典の単語のカバレッジの比較を、和→英→和によって元に戻るかどうかという視点に基づいて行った。その中で、特に英日対訳辞書の充実と日本語の換言語の抽出について述べた。本稿では、英と和を逆にし、英和と和英を「和」で対応を取ることにより、英語の換言語を幅広く抽出することを試みたので、その結果を報告する。

表１：辞典から抽出した語数

	見出し語数(A)	訳語数(B)	B/A
英和	46469	141726	3.0
和英	28395	45937	1.6

2　英和・和英のマッチング

2.1 英和・和英のマッチングの概要

まず、英和辞典と和英辞典のカバレッジを比較するため、英和の和と和英の和とを対応付け、英→和→英の変換で元の英語に戻るかどうかの視点に立ち、以下のように分類した（表２）。ただし、E1 は英和の英語見出し、J1 は英和の日本語訳を表す。

(a) そのままの形で元に戻る（和英に J1 のエントリがあり、英訳として E1 が存在する）
(b) 形態素の１部分に戻る（和英に J1 のエントリがあり、英訳として E1+α が存在する）
(c) 和英辞典に J1 のエントリはあるが、E1 に一致する英訳が見つからない
(d) 和英辞典に J1 が見つからない

表２：英和・和英のマッチング結果

(a) E1→J1：J1→E1	14585 (10.3%)
(b) E1→J1：J1→E1+α	1664 (1.2%)
(c) E1→J1：J1→E2	52462 (37.0%)
(d) E1→J1：　×	73015 (51.5%)
合計	141726 (100.0%)

なお、英和辞典の見出し語１語に複数の訳語が記述されている場合、それらを別カウントした。これらのうち、(c)の場合で、英訳の語数が不足しているために一致しなかった場合は英和の英を取り込むことで英語の表現のバラエティーを増すことができる。ただし、語義が発散したために一致しなかった場合が考えられるので、(c)からランダムに 100 語抽出して、詳しく分析した。

2.2 英が一致しない場合の分析

その結果を（表３）に示す。なお、微妙なものが

多く、はっきりと分類するのが難しいものが少なくなかった。これを見ると、別の表現・広義→狭義・狭義→広義等の違いはあるものの、英→和→英で元に戻らない場合は、ほぼ同じ意味を別の表現の英語で記述したものが多いことがわかる。これらを効率的にグルーピングできれば、英語の換言語が抽出できると思われる。また、注意を要するものとして、ニュアンスがやや異るものがあった。例えば、同じ「やり損なう」に対応する英語でも「blunder」は大へま、「fail」は失敗という違いがある。また、同じ「荒廃」に対応する語でも「decay」は年月による劣化、「devastation」は災害による被害という違いがあり、日本語は同じ語でも、英語は細かく場合分けをする必要があることを示している。これらの英語は全く同じ意味ではないが、日本人にとって混同しやすい語という意味で、抽出することは有意義であると思われる。つまり、日本語でマッチングを取って英和辞典の英語を取り込むことで英語の換言語を抽出すれば語彙数の増加につながる。

表3：英が一致しない原因の分類

原因	度数	例
別の表現	37	bluff －絶壁－ cliff help －救助－ rescue
品詞違い	26	restful －落ち着いた－ feel at ease sneeringly －冷笑して－ sneer at
ニュアンスがやや異る	9	blunder －やり損なう－ fail decay －荒廃－ devastation
広義→狭義	7	past －経歴－ record mate －配偶者－ spouse
まれな用法	5	inner man －胃袋－ stomach nick －刑務所－ prison
表記のゆれ	4	make-up －メーキャップ－ makeup
用例のみ	4	fix －窮地－ drive into a tight corner
狭義→広義	3	write out －除く－ remove
その他	5	

なお、日本語と英語の品詞が違うためにうまく元に戻らなかったものが約4分の1あった。これらは、例えば、

sneeringly →冷笑して：冷笑する→ sneer at

の場合、日本語を「冷笑して→冷笑する」のように和英の見出しに載っている形に変化させてマッチングしたために、英語の品詞が一致しなくなったと言える。

3　英語換言語の抽出方法

表4：英和を用いた英語換言語の収集

	英和	和英
例1 (抽出)	business →職業 calling →職業 career →職業 craft →職業 occupation →職業 profession →職業 …	職業→ occupation 職業→ profession 職業→ trade 職業→ job
例2 (除外)	brush →触れる greet →触れる meet →触れる note →触れる scrape →触れる touch →触れる	1. 触れる→ touch 2. 触れる→ touch on 触れる→ refer to 3. 触れる→ be against

筆者らは、(表4,例1）のように、英和辞典の日本語訳が同一であるエントリを寄せ集め、その英語見出しを用いて英語換言語のグループを作成した。その結果、単に和英辞典の英訳を集めた場合よりも英語の表現を増やすことができた。例えば、マッチング対象の日本語が「職業」の場合を見ると、和英の英だけなら「occupation」「profession」「trade」「job」の4語なのに対し、英和の英を加えると、「business」「calling」「career」等の多くの英語を得ることができた。

ただし、マッチング対象の日本語が広い意味を持つ場合、抽出グループ内に意味のずれた英語が含まれてしまうため、このような場合を抽出対象から除外した。対象外とするかどうかの判定方法は、和英辞典の記述を参照し、英訳が複数の語義番号に分かれている場合を、「日本語が広義である」と判断するという方法を採った。例えば（表4,例2）、「触れる」の場合は、和英辞典の記述が

語義1：(さわる) touch
語義2：(言及する) touch on, refer to
語義3：(抵触する) be against

のように3つの語義に分かれているので、「触れる」によるマッチングは対象外とする。その結果、「note」「scrape」といった意味のずれた英語が1グループ内に拾い出される場合を取り除くことができる。

4　抽出結果

4.1 抽出結果

抽出結果は（表5）の通りである。参考文献[3]で日本語の場合が 1257 組であったのに比べ、6676

組はかなり多いと言える。これは、和英辞典より英和辞典の方が規模が大きいことと、和英辞典は複数の語義番号に分かれているエントリが英和よりも少ないことに起因すると思われる。

表5：英語換言語の抽出結果

抽出換言組数 （除外組数）	6676 組 1143 組
1組あたり語数：和英のみ	1.8 語
1組あたり語数：英和を追加	5.0 語

4.2 評価方法

次に、抽出結果の品質を見るため、抽出した換言語の組からランダムに100組選び、英語ネイティブによる評価を行った。選んだ100組に含まれる英単語の中から、和英の英からランダムに1語(X)、英和によって新しく増えた語からランダムに1語(Y)を選び、意味の妥当性の評価と品詞の一致性の評価を行った。意味の判定は、各XYのペアに対して以下の判定をすることで評価した。

A：XとYの意味が完全に一致する。
B：XよりYの方が語義が広くなった
C：XよりYの方が語義が狭くなった
D：両者の語義の一部分のみに重なりがある
E：ナンセンス

また、品詞については、英語ネイティブが必ずしも品詞体系を完全に理解しているわけではないので、Xを使った例文を思い浮かべ、それをYに置き換えた場合、品詞や修飾関係が文法的に変にならないかどうかの判定をしてもらうことで評価した。その際、意味が変わるかどうかは考えなくてよいとした。すなわち、

0：文法的に変にならない
X：文法的に変になる
?：両方の場合がある

なお、マッチングを取った日本語は伏せ、英語のみによって判定した。

4.3 評価結果

評価結果を（表6）に示す。また、広義の日本語の除外についての性能を見るため、除外した語についても同様の評価を行った（表7）。これを見ると、半数弱がAと評価され、A～Dは80%である。Aのうち X が10ペアあるが、

frivolous / levity ： 軽薄(な)

のように、日本語では「な」の有無などの違いがあるものの実質的には同一語となる場合でも、英語では品詞が違ってくるという場合が多かった。また、英単語を提示された時に辞書のエントリ全てが思い浮かぶわけではない。例えば、

surrender / capitulation ： 降伏

はAXと判定されたが、正しくはA?のはずである。「surrender」を英和辞典で見ると名詞も載っているが、直感的には動詞しか思い付かないためと思われる。

表6：英語換言語の評価結果

判定	ペア数	内訳
A	44	(0=26 ?= 8 X=10)
B	5	(0= 3 ?= 1 X= 1)
C	5	(0= 4 ?= 0 X= 1)
D	26	(0=17 ?= 3 X= 6)
E	20	(0= 7 ?= 2 X=11)
合計	100	(0=57 ?=14 X=29)

表7：除外した語の評価結果

判定	ペア数	内訳
A	20	(0=13 ?= 2 X= 5)
B	9	(0= 6 ?= 0 X= 3)
C	11	(0= 8 ?= 1 X= 2)
D	22	(0= 7 ?= 3 X=12)
E	38	(0=19 ?= 4 X=15)
合計	100	(0=53 ?=10 X=37)

また、Eの20ペアは意外に多いように思われるが、これについては次章で考察する。

一方、（表7）の除外語のうち、Aが20ペア存在するが、除外対象といえども意味のずれたペアばかりになるわけではなく、語義区分によっていくつかに細分されるというのが実際のところであろう。したがって、除外した組を何らかの方法で細分できれば、さらに多くの換言語を抽出できると思われる。ただ、英和の和は日本語訳が載っているだけで、和英の何番目の語義に対応するかという記述はないので、何らかの工夫が必要である。

5 考察

5.1 換言語のE判定の語の分析

（表6）のE判定が20ペアあった原因を調べる

ため、各ペアのY（英和辞典のエントリ）を個々に詳しく見てみると、あまり使われない用法が多いことがわかった。くだけた日常語であることを表す記号が付いていたエントリが4ペア、英和辞典が後の方の語義区分になっていて普通使われない用法と思われるものが8ペアあった。例えば、

toilet / cloakroom ： トイレ

の場合、cloakroom を辞書で引くと確かにトイレの意味が載っているが、旅先で cloakroom の場所を尋ねたら、おそらく、トイレではなくクロークに案内されるであろう。それゆえ評価者は E と判定したと思われるが、E の判定でも必ずしも抽出誤りではないことがわかる。

また、英和辞典の記述に疑問がある場合も2ペアあった。例えば、

syringe / squirt ： 注射器

の場合、英和辞典には squirt に注射器の意味が載っているが、英英辞典には注射器の意味は載っていない。

5.2 後方の語義の切り捨てによる精度改良

もし、換言語を幅広く集めることよりも精度を重視したい場合、前節で述べたような普通使われない用法を除去することで抽出精度を改善できる可能性がある。一般に英和辞典は使用頻度の高い用法ほど最初に記述されているため、語義番号の後の方を落とすとよい。これを見るために、（表6）を英和辞典の語義番号別に分類し、判定結果を調べた（表8）。これを見ると、語義番号が小さいエントリから抽出した英語ほど判定結果が良いことがわかる。判定 A のものを抽出したい場合は語義1と2でおおむね網羅されることを示している。ただ、語義番号別に分けると個数が少なくなってしまい、後の方の語義番号は精度が悪いことを確かめるためには、調査数を増やす必要がある。

表8 ：英和の語義番号別の評価結果

語義番号	A～E計	判定A	判定A～D
語義1	57	34 (60%)	50 (88%)
語義2	21	8 (38%)	13 (62%)
語義3	12	2 (17%)	9 (75%)
語義4	2	0 (0%)	2 (100%)
語義5	3	0 (0%)	3 (100%)
語義6以降	5	0 (0%)	3 (60%)
合計	100	44	80

5.3 WordNet との比較

英語のシソーラス辞書である WordNet[4]と比較するため、「職業」に対応する英語を含んだエントリを幾つか WordNet から探し、（表9）に挙げた。これを見ると、WordNet は組が細かく親子関係に分かれているのに対し、本稿の方法は、profession, work 等の、違いがあるのかどうか英語ネイティブでないと直感的に分かりにくい語が同じ組に挙がっている。このような組を使えば、文のマッチングを取る場合の再現率を向上させるのに効果的である。

また、walk, way のように語彙に曖昧性のある語と business のように曖昧性の少ない語が挙がっていることから、曖昧性のある語を曖昧性の少ない語に置き換えて文のマッチングを取るという用途にも応用できることを示している。

表9 ：WordNetとの比較例

WordNet	occupation / business / line of work / line
	├ profession
	├ trade / craft
	└ job / employment / work
提案手法:「職業」	和英:occupation / profession / trade / job 英和:business / calling / career / craft / game / line / occupation / profession / pursuit / trade / vocation / walk / way / work

6 まとめ

本稿では、英和辞典と和英辞典のエントリを日本語でマッチングを取ることにより、英語の換言語を幅広く抽出する方法を報告した。また、抽出結果の英語ネイティブによる評価を行った。さらに、精度の調整をしたい場合は語義番号を用いてできることを示した。本稿の結果は文のマッチングを取る用途に有用であると思われる。特に、違いがあるのかどうか直感的に分かりにくいような語を含んだ文のマッチングを取りたい場合に効果を発揮すると考えられる。

参考文献

[1]ｽｰﾊﾟｰｱﾝｶｰ英和辞典ﾃﾞｰﾀﾍﾞｰｽ, 学習研究社(1999)
[2]ﾆｭｰｱﾝｶｰ和英辞典ﾃﾞｰﾀﾍﾞｰｽ, 学習研究社(1995)
[3]鷹尾・下畑・今村・柏岡: 和英／英和辞典のカバレッジの比較とその応用, 言語処理学会第7回年次大会, pp.58-61(2001)
[4]WordNet 1.6, Princeton University (1997), http://www.cogsci.princeton.edu/~wn/