和英／英和辞典のカバレッジの比較とその応用

鷹尾 和享　　下畑 光夫　　今村 賢治　　柏岡 秀紀
ATR音声言語通信研究所
E-mail: {kazutaka.takao, mitsuo.shimohata, kenji.imamura, hideki.kashioka}@slt.atr.co.jp

1　はじめに

ＡＴＲ音声言語通信研究所では話し言葉の機械翻訳システムＴＤＭＴを作成しているが、その過程で、市販の和英辞典・英和辞典の電子データを利用する機会を得た。そこで筆者らは、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。本稿では、まず和英の英が英和に見つからない場合等の分類を行い、カバレッジについての考察を述べる。さらに、それぞれの場合についての特徴を分析し、そこから得られる情報の用途について、翻訳システムの辞書への利用等の観点で考察を行う。その際に、多くの有用な情報が得られることがわかった。特に、英語の複合語が一語の日本語訳になる場合や、日本語の換言を抽出できることがわかったので、それについての議論を深める。

2　カバレッジ比較の概要

筆者らは、学研のニューアンカー和英辞典[1]とスーパーアンカー英和辞典[2]の電子データを用い、機械翻訳用の対訳ペアを抽出した。抽出した語数は表１の通りである。見出し語数には小見出しの派生語や複合語も含んでいる。対訳ペア数とは、見出し語１語に複数の訳語が記述されている場合、それらを別ペアとしてカウントした。

表１：辞典から抽出した語数

	見出し語数	対訳ペア数
和英	28395	45934
英和	46469	141000

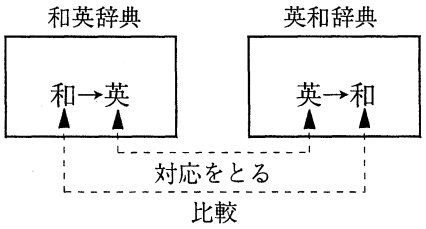

図１：和→英→和

次に、図１のように、和英の英訳と、英和の英語見出しとの対応をとることにより、和→英→和で最初の「和」と最後の「和」とがどの程度元に戻るかの分析を行い、表２のように分類した。次章では各場合についての分析・考察を行う。

表２：和→英→和で対応をとった結果

復元	そのままの形で元に戻る	12717（27.7%）
可能	形態素の１部分に戻る	2589（ 5.6%）
復元	ｴﾝﾄﾘはあるが和が一致しない	14088（30.7%）
不可能	和→英の英が英→和にない	16540（36.0%）
合計		45934(100.0%)

3　対応結果の分析・考察

3.1 そのままの形で元に戻る場合

和→英→和でそのままの形で元に戻る語は、表現のゆれが少なく、語義の曖昧性の少ない語が抽出されていることがわかる（表３）。さらに、戻った和が英和辞典の最初に記述されている訳語かどうかで分類すると、最初の訳の場合はそのまま意味が伝わる語であり、２番目以降の訳語の場合は多少曖昧性が生じていることがわかる。このことは日英翻訳結果の曖昧性の検証に利用できると思われる。

表3：そのままの形で元に戻る語の例

英和の和が最初の訳語	ｱｲｽｸﾘｰﾑ－ ice cream －ｱｲｽｸﾘｰﾑ、 重い－ heavy －重い、 足－ leg －足
英和の和が2番目以降の訳語	国民－ people －人々／国民、 アース－ earth －地球／大地／アース、 新しい－ fresh －新鮮な／さわやかな／新しい

3.2 形態素の1部分に戻る場合

和英和で形態素の1部分に戻る場合は、英和の和訳を形態素に分け、そのうちの1形態素に戻るものである。それを品詞パターンで分類し、表4に多い順に10パターンを挙げた。

これを見ると、「安全」「以前」「屋内」のように、英語が形容詞・副詞である場合は、和の不一致の原因は単に助詞類の有無であるものが多いことがわかる。また、「削除」「制限」のように、英語が動詞である場合は、サ変名詞とサ変動詞の違いや、「制限を設ける」のように名詞と対で使われる動詞の有無が原因であるものが多い。

表4：形態素の1部分のうち多いパターン

パターン	度数	例
ADJ:X+な	434	安全－safe－安全+な
ADJ:X+の	301	以前－former－以前+の
CN:X+名	213	アルプス－Alps－アルプス+山脈 受付－receptionist－受付+係
CN:名+X	176	期限－deadline－最終+期限 相手－match－競争+相手
V:X+補動	126	削除－delete－削除+する
CN:名+の+X	73	消印－postmark－郵便+の+消印 車掌－conductress－女性+の+車掌
CN:X	66	希望する－hope－希望
CN:X+の+名	49	生まれ故郷－hometown－生まれ故郷+の+町 あひる－drake－あひる+の+雄 ぶどう－vine－ぶどう+の+木
V:X+を+動	44	居眠り－doze－居眠り+を+する 制限－limit－制限+を+設ける
ADV:X+に	40	屋内－indoors－屋内+に

英語：ADJ=形容詞、CN=名詞、V=動詞、ADV=副詞
日本語：X=一致する日本語形態素、名=名詞類、
補動=補助動詞、動=本動詞

一方、「アルプス」「受付」のように、英語が名詞の場合は、日本語訳では別の名詞が付いているものが多い。たとえば、「アルプス」だけで通常は山脈を意味することがわかるので、「山脈」は説明のわかりやすさのために付いていると言える。一方、「受付」だけなら場所と担当者の両方の意味が考えられるが、「係」が付加することによって語義が狭まっていることがわかる。前者は英日翻訳での長短2種類の訳し分けに利用できる。また、後者は日英翻訳の日本語入力文の曖昧性をチェックするのに利用できると考えられる。

3.3 エントリはあるが和が一致しない場合

表5：和が一致しない原因の分類

原因	度数	例
別の表現方法	52	暑苦しい－stuffy－風通しの悪い 数数－lot of－たくさん 療養－recuperation－静養
狭義→広義	15	勢い－power－力 一行－group－群れ/集団 改修する－repair－修理する
かな書き／漢字異表記	11	搾る－wring－しぼる 闘う－fight－戦う
品詞の違い	7	根－naturally－生まれ付き 輪番－alternately－交互に/かわるがわる
広義→狭義	6	腕－ability－能力/手腕 海老－spiny lobster－イセエビ
ニュアンスがやや異る	6	意見－advice－助言 気の長い－long-range－長距離の/長期の
自動詞・他動詞の違い	1	栄転する－promoted－昇進させる
カバレッジの違い	1	早生－early－早い/時期より早い
その他	1	

和英と英和でマッチする英語のエントリはあるが、最初と最後の和が一致しない場合について、ランダムに100語を取り出し、不一致の原因を分析した（表5）。

これを見ると、「暑苦しい→風通しの悪い」のように、約半数が表現方法の違いによる不一致であり、換言語句として利用可能であると思われる。また、言語文化に関連する原因としては、「勢い

→力」のように、和英和で戻るときに狭義から広義になったもの、逆に「腕→能力／手腕」のように広義から狭義になったもの、「意見→助言」のようにニュアンスがやや異るものがあった。また、文法に関連する原因としては、「根(名詞)→naturally(副詞)」のような品詞の違い、「栄転する－ promote」のような自動詞／他動詞の違いによる不一致があった。また、「搾る→しぼる」のように、かな／漢字の違いのために、本来なら3.1に分類されるべきものが、不一致になっている場合も見られた。

3.4 和英の英が英和にない場合

和英の英が英和にない場合についても、ランダムに 100 語を取り出し、その原因を分類した(表6)。

これを見ると、日本語1語が英語の複合語に対応する場合がおよそ半数を占めることがわかる。これらの語の日と英を逆にし、英日機械翻訳に取り入れると、英語の複合語が1語の日本語に翻訳できるものを抽出したことになる。英和辞典の英と和を逆にして日英機械翻訳に利用するアイデアは白井ら[3]が提案しているが、同様に、和英辞典を英日機械翻訳に利用すると有用である。これについては次項で詳しく述べる。

次いで、日本語・英語とも複合語の場合もかなりあることがわかる。これらも日と英を逆にして英日機械翻訳に利用すると、辞書の規模を効果的に拡大できると思われる。

また、日本語が成句調のエントリがあった。たとえば、「微力+ながら」「騒が+せる」のように、特定の機能語がセットで使われる場合は特定の英訳になることを示している。これらを単に対訳辞書に取り入れるだけで済ませることは難しいので、ルールの整備が必要と思われる。

英訳が説明調のものも見られたが、これらは長短の訳し分けに利用できると思われる。たとえば「駅弁」は駅の案内所での会話であれば「box lunch」で十分であるが、ホテルのフロントならば長い訳語の方が適切に意味を伝えられると思われる。

日本文化特有の語の音訳も見られた。これらの語は英和辞典には普通載せないと思われるが、実際の会話で日本人と外国人が話すような場合、外国人側から「tabi」等の発話をすることがあると思われるので、英日機械翻訳の辞書に必要であると思われる。

表6：英がない場合の分類

原因	度数	例
英語の複合語が1語の日本語	49	がらがら－ almost empty 群島－ group of islands 習う－ take lessons 要職－ important post 野党－ opposition party
複合語同士	21	税関申告書－ customs declaration form 特別料金－ extra charge
日本語が成句調	9	微力－ though i'm afraid won't much 騒がせる－ caused sensation
英語が説明調	7	駅弁－ box lunch sold at a railroad station 点字ブロック－ raised stop line on the platform for the blinds
1語同士	4	過保護－ overprotection 肥満－ fatness
日本文化の音訳	4	足袋－ pair of tabi 立秋－ risshuu
品詞の違い	2	刈り入れ－ harvesting ヒッチハイク－ hitchhiking
その他	4	

4 応用例

4.1 英日機械翻訳への利用

前項の表6で、英語の複合語が1語の日本語になる場合を示したが、さらに詳細に考察することで、英日機械翻訳の改良の手がかりを得ることができる。一般に、機械翻訳は直訳調になりがちであるが、和英辞典の日本語見出しにある語を英日機械翻訳の日本語訳として生成すると、より平易な訳が得られる。

まず、「要職」「野党」のように、英語の形容詞＋名詞が、日本語では漢字を組み合わせた熟語に対応することがわかる。もし「要職」が和英辞典

に載っていなければ、漢字の意味を理解して「重要な職務」を辞書で引いて英語に直すわけであるが、この逆のプロセスを経れば、英語の形容詞＋名詞から日本語の熟語を生成し、短い日本語訳を得ることができると思われる。その際、たとえばpartyでは、

opposition party －野党
farewell party －歓送会
wintering party －越冬隊

のように、漢字が「党」「会」「隊」の複数考えられる場合もあるので、訳し分けの工夫が必要と思われる。

同様に、「習う－take lessons」のように英語の動詞＋目的語も日本語では1語の動詞にできる語があることを示している。あるいは、「産卵－lay eggs」のように、漢字の熟語にできる場合もある。

また、日本語は擬態語が豊かであるという特徴がある[4]。ＴＤＭＴの英日翻訳の日本語訳には擬態語が比較的少ないが、表6の日英を逆にすると「がらがら」のような擬態語の生成ができることになり、より話し言葉らしい日本語訳になると思われる。

4.2 日本語から日本語への換言の抽出

3.3 で約半数が換言語句に利用できることを述べたが、さらに精度の高い抽出方法を検討することにする。単純に和英の英が一致する語を集めると、たとえば「力／能力／権力／強国／電力－power」のような、英語が広義であるようなペアが混入してしまう。また、英和の和訳の表現のバラエティーは限られた個数しか載っていない場合が多い。そこで、英和を利用し、和訳の語義が1つのみの英語を「語義のゆらぎがない」と見なし、その英語に和英の英訳が一致するような語をペアとして抽出し、それらの日本語見出しを取り出すと、精度の高い換言が得られる。

抽出できたペア数は 1257 であり、表7に例を挙げる。これによって、英和辞典に載っている日本語訳語に加え、和英辞典の日本語見出しから換言を得ることができた。たとえば、「abolish」は英和辞典の日本語訳では「廃止する」のみであるのに加え、和英辞典の日本語見出しから「全廃する」「撤廃する」の表現を得ることができた。なお、大部分はうまく取り出せたが、注意を要するものが若干あった。たとえば、「女優／俳優／役者」のように性別を表す語と表さない語が混ざっているもの、big dipper のように和英と英和のカバレッジの違いに起因するゴミ等が見られた。

表7：日本語から日本語への換言の抽出例

うまく取り出せた例	全廃する/撤廃する/廃止する－abolish 空港/飛行場－airport ブティック/洋品店－boutique
注意を要する例	女優/俳優/役者－actress ジェットコースター/北斗七星－big dipper

5 まとめ

本稿では、和英辞典と英和辞典の比較を和→英→和で元に戻るかという観点で分析を行い、その結果多くの興味深い情報が得られることを示した。特に、英語の複合語が1語の日本語に翻訳できるものを抽出し、英日機械翻訳に利用すると効果的に改良できる可能性があることを示した。今後はこれらの情報を機械翻訳システムＴＤＭＴに活用していきたい。

また、和英辞典で英訳の単語の記述がなく、用例文だけが書いてあるエントリが少なからず見受けられ、対訳を正しく抽出するのが困難なエントリがあった。日本語に対する英語の訳語の特定がもともと難しい場合もあるが、それらを機械翻訳用の辞書に適した形で取り出すことについてはさらに検討の余地がある。

参考文献

[1]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)
[2]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)
[3]白井諭他: 英単語に対する述語性の連語的日本語訳語の分析, 情報処理学会研究報告, 97-NL-122-2（自然言語処理研究会）, pp.7-12(1997)
[4]金田一春彦他: 学研国語大辞典第二版, pp.2137-2138, 学習研究社(1978)