日英機械翻訳のための大規模慣用表現辞書の構築

田村 真子　亀井 真一郎　土井 伸一　山端 潔
NEC 情報メディア研究所

1. はじめに

複数の単語が組み合わさって句単位で特別な意味を生じる表現(以下慣用表現と呼ぶ)を適切に処理することは、機械翻訳の質を向上するための課題の一つである。

従来、慣用表現に関して言語学的な観点[1, 2]だけではなく工学的な観点からも多くの研究がなされている[3, 4, 5]。また、慣用表現の言語現象の分析や処理方法の提案に加えて、最近ではコンピュータを用いた自動抽出などによる慣用表現の収集の試みもある[6]。しかしこれら従来の研究では、日本語として振る舞いが特殊な表現に焦点が当てられており、翻訳の質の向上の鍵を握る英語表現との関係まで十分言及されているものは殆んどない。

今回我々は日本語見出し数で約20,000の用言句相当慣用表現を収集し、英語訳を付与して対訳辞書を作成した。我々が慣用表現として収集の対象とした表現には、従来研究で慣用表現と呼ばれている表現の他に、特に日英翻訳という観点から日本語では普通の表現だが英語では一語になるなど対応が取りにくいものも含めた。

本稿では、日英機械翻訳のための慣用表現の収集法について述べる。

2. 機械翻訳における慣用表現

今回我々は、句単位で特別な意味を生じる、という従来から言われている慣用表現の特徴を、機械翻訳の観点から句単位の表現に対して特定の訳語が決まる、という意味に拡大解釈し、以下の4種類を広く日英慣用表現と扱うことにした。

1. いわゆる慣用句
   (例)「さぼる」の意味の「油を売る」
   = to idle away one's time
2. 機能動詞表現
   (例)「電話する」の意味の「電話をかける」
   = to telephone
3. 部分的に訳語が特定のものに決まる表現
   (例)「湯を沸かす」 = to boil water
4. 日英の対応が一対一ではない表現
   (例)「頭がいい」 = clever
   (例)「ペンキを塗る」 = paint

1. は従来研究でいわゆる慣用句と呼ばれているものであり、句全体で、構成語それぞれの意味を足したものではない別の意味を生じる表現である。2. は従来研究で慣用表現の一つとして個別の研究もなされており[2]、体言部分が句全体の意味を担っていることが特徴の表現である。3. は複数の日本語単語の組み合わせに対して、そのうちのいずれかの単語に対応する英訳語が特定のものに決まる表現である。2. と3. はまとめてコロケーションと呼ばれることもある。4. は日本語も英語も従来の定義からすると慣用表現とは言えないが、複数の日本語単語の組み合わせに対して一語の英訳語が対応するため、翻訳の際には句単位で処理する必要のある表現である。今回、我々は4. のような句単位の日英対訳表現も慣用表現として収集することにした。

また慣用表現には、全体で名詞句として振舞うもの、副詞句として振舞うもの、動詞句あるいは形容詞句の用言句として振舞うものなどがあるが、今回我々は用言句相当慣を収集の対象とした。用言句は文の中心となることが多く、用言句として振舞う慣用表現を適切に処理することは文全体を適切に処理することの大きな一歩であるからである。

今回我々は、上記1. ～4. の日英慣用表現対を合計約20,000表現収集し、それらに英訳語を付与した。以下にそれらの収集と英訳付与について述べる。

3. 日本語慣用表現の収集

慣用表現の大きな特徴の一つとして、慣用表現を構成する語の組み合わせは"慣用的に"決まるという語彙の特徴がある。その"慣用的に"決まる特定の語の組み合わせを個別に処理するためには、慣用表現を大量に収集し辞書を構築する必要がある。"慣用的に"決まる特定の語の組み合わせを収集するには人の判断が必要であるが、大量の表現を効率的に収集するためにはやはりコンピュータの助けを借りる必要がある。そこで、今回我々は、まずコンピュータを用いて大まかなデータを作成し、そのデータの中から人が判断して慣用表現を収集した。

Collecting of Verbal Idiomatic Expressions and Development of a Large Dictionary for Japanese-to-English Machine Translation
TAMURA Shinko, KAMEI Shin-ichiro, DOI Shin-ich, YAMABANA Kiyoshi
NEC Information Technology Research Laboratories.

3.1. テキストデータからの収集

収集の作業データを作成するために、まず我々はどのような語が慣用表現の構成語として現れやすいかの傾向を調べた。

今回収集の対象とした用言句相当慣用表現は一般に以下のような形式を持つ。

体言 + 格助詞 + 用言 (+ 付属語)

我々は、上記の形式を持ち、いわゆる慣用句として従来研究 [1, 2, 7] で言及されている約 1,800 の用言句相当表現について、それぞれの体言部と用言部にどのような語彙が現れるかの前分析を行なった。

その結果、用言句相当慣用表現の体言部に現れていた語数 (異なり数) の約 4 割が「手」のような身体の部分を表す身体名詞であり、約 3 割が「攻撃 (する)」の「攻撃」のような動作や行為を表す動作性名詞であった。残りの約 3 割は、「空」などの自然を表す名詞、「埒があかない」の「埒」などそれ単独では殆んど用いられないような名詞であった。

用言部では、そこに現れるものの 9 割以上が「上げる」のような和語動詞であった。日英機械翻訳用辞書で、日本語用言の種類の分布を採ったところ、その 5 割以上が「連絡する」のようなサ変動詞をはじめとする漢語であったので、慣用表現に現れる用言の分布は、一般の用言のそれとは全く異なっていることが分かった。

上記の結果を元にして、我々は身体名詞と和語動詞を中心とした約 1,700 語を、慣用表現を構成する傾向にある語と考え今回の収集のキーとした。収集はまずそれらのキーを含む文をキーの直前・直後の 5 文字でユニークソートした約 720,000 文をプログラムを用いて用意し、それらの中から慣用表現の意味分布図 [5] などを参考にして、人が慣用表現を抽出するという作業を行なった。この作業で約 8,000 件の日本語慣用表現を得た。表 1 に収集キーの種類と収集した表現の例を示す。

表 1：収集キーの種類と収集した表現の例

		キー	収集した表現の例	
キーの種類	用言	あげる	経をあげる	ネをあげる
			気勢をあげる	男をあげる
			軍配を上げる	棚に上げる
		つく	はずみがつく	差がつく
			見通しがつく	気がつく
		切る	トップを切る	電源を切る
			領収書を切る	しらを切る
	身体名詞	手	手を焼く	手を切る
			手に入る	手を出す
		口	口を出す	口を切る
			口が軽い	口に出す

3.2. 機能動詞表現の収集

日本語では「連絡する」の意味で「連絡をとる」というように 1 語で言うのとほぼ同じ意味のことを句の形で言うことがある。「連絡をとる」では名詞「連絡」が句全体の意味を担っており、「とる」は「何かを物理的につかむ」という意味ではなく、動詞活用語尾「する」と同等なものとして補助的に機能している。また、この種の表現では動詞部分が「〜られる」のような助動詞や「〜できる」のような補助動詞と言い替えられる場合も多い。例えば、「期待がかかる」は「期待される」と言い替えることができる。このような表現は新聞などでも頻繁に用いられる慣用表現の一種であり、機能動詞表現と呼ばれて個別に研究もされている [2]。前述の動作性名詞はしばしばこのような機能動詞表現を構成する。動作性名詞には、1)「連絡 (する)」などのサ変名詞、2)「叫び」など和語動詞の派生名詞形 (=連用形)、3)「会議」などの動作・行為を表す普通名詞がある。また、動作性名詞と共に機能動詞表現を構成する動詞は機能動詞と呼ばれるが、それらはこれまで我々が見てきた限り全て和語動詞である。機能動詞と成り得る和語動詞の一例を表 2 に示す。

表 2：機能動詞の例

和語動詞	対応する助動詞など	機能動詞表現の例
取る	する	連絡を取る
かかる	〜れる	期待がかかる
あう	〜れる	いじめにあう
いく	できる	合点がいく

我々は、前述の動作性名詞と表 2 に示すようなそれと結び付きやすい和語動詞をキーとして機能動詞表現の収集を個別に行なった。

収集ではまず、動作性名詞のリストと、それらと共に機能動詞表現を構成すると思われる和語動詞 (= 機能動詞) のリストを用意した。今回は、日英機械翻訳用辞書にあるサ変名詞 (= 漢語動詞の語幹) 約 17,500 語と和語動詞の連用形約 10,500 語の約 28,000 語を動作性名詞のリストとした。機能動詞としては、和語動詞約 100 語を従来文献 [2] を参考にしてリストアップし、更にそれらを表 2 の例が示すように機能動詞表現においてそれぞれが対応する活用語尾 / 助動詞 / 補助動詞の種類で分類した。例えば、「連絡を取る」は「連絡する」に言い替えることができ、この場合「取る」は「連絡する」の「する」に相当する機能を果たしている。我々はこのようにサ変動詞活用語尾「する」に相当する機能動詞の種類を「基本」と呼んでいる。また、「期待がかかる」は「期待される」に言い替えることができ、この場合「かかる」は「される」という受身を表す

助動詞に相当する。「(さ) れる」や「(ら) れる」に相当する機能動詞は「受身」と呼んでいる。更に、「納得がいく」は「納得できる」に言い替えることができ、この場合は「いく」は「できる」という可能を表す補助動詞に相当するので、「可能」の機能動詞と呼んでいる。作業者は、上記2つのリストを照らし合わせながら、動作性名詞と機能動詞に成り得る和語動詞を組み合わせて、1) 名詞が句全体の意味を担っており、2) 動詞は助動詞に置き換えられるなどの補助的な働きしかしていないものを機能動詞表現としてリストアップする。今回我々はこの方法で機能動詞表現を約12,000表現収集した。

(収集した機能動詞表現の例)
期待をかける (≈期待する)
蹴りを入れる (≈蹴る)
承認を受ける (≈承認される)
納得がいく (≈納得できる)

4. 英訳語の付与

今回の対訳データ収集は主に以下の3つの方法で行なった。

1. 日本語の言い替えを利用した対訳付与
2. 機械翻訳用英日辞書からの対訳データの抽出
3. 人手による対訳付与

以下に1. ～ 3. の方法をより詳細に述べる。

4.1. 言い替えを利用した対訳付与

我々が収集した慣用表現のうち、機能動詞表現では、前述したように動作性名詞が句全体の意味を担っており、動詞はその動詞としての意味ではなく動詞活用語尾「する」などとして補助的に機能している。従って機能動詞表現に対する訳語はそれを構成する動作性名詞の動詞用法に対する訳語と基本的に同じでよいと仮定できる。例えば、「連絡をとる」に対しては「連絡する」と同じ訳語が考えられる。そこで我々は、この特徴を用いて機能動詞表現に対訳を付与することにした。

まず、機能動詞表現を言い替えた先の動詞 (サ変動詞または和語動詞) の英訳語を我々の既存の機械翻訳用辞書から取り出し、その動詞の訳語と機能動詞表現との実際の日英対訳の妥当性を英語の分かる日本語ネイティブスピーカが判断した。これは、言い替えた先の動詞に対する英訳語が必ずしも機能動詞表現に対しても適切であるとは限らないからである。例えば、「電話を受ける」の「受ける」は機能動詞リストで受身の意味を持つものとして分類されているが、この表現に対応する適切な英語表現は「telephone」を受動態にした「be telephoned」ではない。更に日本語の分かる英語ネイティブスピーカが訳語の英語としてのチェックを行なった。

この方法により我々は、収集した機能動詞表現のうちの約7,000表現に対する英訳語を得た。

4.2. 英日辞書を用いた日英対訳の収集

2番目の対訳収集法は、逆方向の辞書すなわち英日辞書を用いた方法である。

従来の機械翻訳の課題として、原言語と目的言語の単語が一対一に対応しないものをどのように扱うかがある。例えば、第2節の4. で挙げた「ペンキを塗る」とその英訳の「paint」は複数の日本語の組み合わせに対して一語の英語が対応している。この種の表現は句全体が一語の英語に対応しているという点で機能動詞表現と共通している。表3は、日本語の体言部の品詞と動詞用法の有無という観点からこのような「日本語: 体言+付属語+用言→英語: 用言」型の日英対訳を分類した表である。

表3で (A) と (B) の日本語の表現は機能動詞表現である。(C) と (D) は日本語が従来の研究の観点から見ると慣用表現ではないので、前節までで述べてきた収集ではこれらの表現は殆んど拾えていない。しかし、日英翻訳という観点から見ると、(C) も (D) も句単位で特別に扱う必要のある表現であり、決して無視できない表現である。今回我々は以下に述べる特徴を手掛かりにして (C) タイプの対訳を収集する方法を考案した。(C) タイプの日英対訳の関係をより詳細に見ると図1のようになる。

図1：(C) タイプの日英対訳の関係

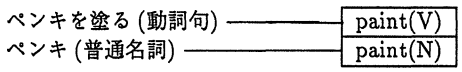

表3：体言部の品詞と動詞用法の有無から見た分類

	日本語 ↔ 英語の例	日本語体言部の品詞	日本語体言部の動詞用法	英語と同一表層の名詞
(A)	電話をかける ↔telephone(動詞)	サ変語幹	有り (電話する)	有り (telephone)
(B)	疑いをかける ↔doubt(動詞)	和語動詞の派生名詞	有り (疑う)	有り (doubt)
(C)	ペンキを塗る ↔paint(動詞)	普通名詞	無し	有り (paint)
(D)	背が高い ↔tall(形容詞)	普通名詞	無し	無し

図1が示す関係から、一つの英語表層に対して名詞と動詞の用法があり、かつ、名詞用法の英語に対応する日本語が普通名詞のとき、英語の動詞用法に対応する日本語には「その名詞＋付属語＋用言」の句があると予測できる。例えば上記の例で、「paint」には名詞と動詞の用法があるが、名詞の「paint」に対応する日本語の「ペンキ」は普通名詞である。この場合、動詞の「paint」に対する日本語には「ペンキを塗る」のような「ペンキ＋付属語＋用言」の句があり得る。今回、この予想をもとにして以下のような手順で(C)タイプの対訳を機械翻訳用英日辞書から取り出した。

1. 機械翻訳用英日辞書のエントリのうち、ひとつの英語表層に対して名詞と動詞の両方の用法があるものを対訳の形で取り出す。
   (例)label →ラベル、label →ラベルを貼る、
   label →ラベルを付けて分類する
2. 1. で取り出した対訳の中で日本語が「名詞＋付属語＋用言」の形式であり、かつ、日英方向の翻訳の観点から見て慣用表現として扱うべきと判断したものを取り出す。
   (例1) ラベルを貼る→ label(対訳として採用)
   (例2) ラベルを付けて分類する→ label
   (説明的なので日英対訳としては不採用)
3. 日本語の対訳が一語のものについては対応する日本語慣用表現を考える。

今回、上記の1. に該当する約3,700エントリ分の英日対訳から、約3,000対の日英慣用表現を得た。

4.3. 人手による対訳付与

第2節の1. のいわゆる慣用句では、句全体で構成語の意味の足し算ではない別の意味が生じる。従って訳語も、例えば「さぼる」の意味での「油を売る」に対しては「idle away one's time」のように、句全体に対して特定の訳語が決まる。また、第2節の3. の表現では、特定の語が組み合わさった場合に限り部分的に訳語が決まる。例えば、「湯を飲む」では「drink *hot water*」だが、「湯を沸かす」では「boil *water*」となる。従って、このような表現に対しては人手による対訳付与が必要である。今回、英語が分かる日本語ネイティブスピーカが対訳付与を行ない、付与された英語のチェックは日本語が分かる英語ネイティブスピーカが行なった。これにより、我々は約10,000の日本語慣用表現に対する英訳語を得た。

また同様の方法で、機能動詞表現と日英対訳が一対一ではない表現に対して、「動詞＋名詞」の訳語の付与を行なった。前述の4.1と4.2の方法により得られたそれぞれの日本語慣用表現に対する訳語は一語の動詞または形容詞である。しかし、機能動詞表現と日英対訳が一対一ではない表現は、統語的な振舞いが一般用言句と殆んど同じであり、翻訳の際に訳語が一語の動詞や形容詞では不都合な場合がある。例えば、「電話する」の意味の「電話をかける」ではその構成語の「電話」に連体修飾語がかかって「重要な電話をかける」のように言うことができるが、このとき、4.1の方法で付与した動詞の「telephone」ではその連体修飾語に対応する適切な翻訳結果が得られない。従ってこのような表現には、対応する英語表現として「make a phone call」のような「動詞＋名詞」の形式を持つ訳語が必要である。そこで、このように日本語の名詞に修飾語が掛かることがありその修飾語が英語側でも連体修飾語として訳されるべき表現に対して「動詞＋名詞」の形式の訳語も付与した。

5. おわりに

本稿では、日英翻訳の観点から用言句相当慣用表現を整理し直し、約20,000の日本語慣用表現とそれに対応する英語表現を収集した方法を述べた。

更に我々は、慣用表現の日英辞書記述モデル[8]を参照しながら、収集した対訳表現に日英翻訳情報を付与し、日英構文変換用辞書を構築した。その辞書は、我々が開発したFEP型英文作成支援ツール[9]に実装している。現在は、本辞書を用いた機械翻訳における質の向上について評価中である。

今後も機械翻訳の質をより向上させるために、句単位で振舞う日英対訳表現に注目し、日本語と英語のずれを吸収できるより良い辞書の構築を目指す。

参考文献

[1] 宮地裕:『慣用句の意味と用法』, 明治書院,1982.
[2] 村木新次郎:" 日本語の機能動詞表現をめぐって " 国立国語研究所報告6 5,1980.
[3] 奥雅博:" 日本語慣用表現の分析と日英翻訳への適用 ", 情処研究会資料 87-NL-62-2, 1987.
[4] 首藤公昭他:" 日本語の慣用的表現について ", 情処研究会資料 87-NL-66-1, 1987.
[5] 亀井:" 日本語の用言句相当慣用表現の分類とその応用 ", 第47回情処全大.
[6] 新納浩幸他:" 語義の特異性を利用した慣用表現の自動抽出 ", 情処学会論文誌 Vol.36 No.8,1995.
[7] 倉持保男他編:" 必携慣用句辞典 ", 三省堂,1982.
[8] 田村他:" 用言句相当慣用表現の日英対訳の型の分類とその応用 ", 第50回情処全大.
[9] Yamabana,K.,Doi,S.,Kamei,S.,Satoh,K., Tamura,S.,Ando,S.:"Interactive Machine-Aided Translation Reconsidered － Interactive Disambiguation in TWP －",*Proc. of NLPRS'95*,pp.368-376.