日本語動詞句相当慣用表現の収集と評価

土井 伸一, 田村 真子, 亀井 真一郎

NEC C&C メディア研究所

1. はじめに

機械翻訳システムは現在、低価格なソフトが売り上げを伸ばすなど様々な形で実用化され、実際に使用される場面も増えてきているが、適切な訳語の選択は依然として大きな課題である。品質向上のためには、各単語ごとの処理を充実するだけでなく単語間の共起・依存関係を考慮する必要があり、特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たり特別な訳語に対応する表現(以下本稿では、「慣用表現」という用語をこの意味で用いる)を適切に処理しなければならない。

日本語のこの種の慣用表現に関してはこれまでに、言語学的、工学的な様々な分析、処理法の研究が行われており、また最近では、コーパスからデータを自動抽出する手法も提案されている[1,2,3,4,5,6,7]。しかしこれらの研究では、主に日本語としての振舞いが特殊な表現に焦点が当てられており、訳語選択のための英語表現との関係まで十分言及されているものはほとんどない。

そこで筆者らは、用言句相当の慣用表現に関して、意味的な分析を行って分布図を作成し[8]、これに従って、日英機械翻訳の品質向上を目的とした大規模慣用表現辞書を構築した[9]。これまでに、日本語見出し数で約3万の日英対訳慣用表現辞書を構築している。また筆者らは、FEP型の英文作成支援ツール「英文名文メイキング」を開発し、構築した慣用表現辞書の一部1万5千表現を実装している[10]。本稿では、この慣用表現辞書の収集・構築方法の概要を述べるとともに、特に体言+格助詞+動詞という形をした動詞句相当表現について17万文の新聞文[11]を対象としてカバレッジ評価を行った結果と、英文作成支援ツールにおいて実現した訳語選択のインタフェース[12]に関して報告する。

2. 日英翻訳の観点から見た日本語慣用表現

日本語慣用表現としては従来は、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、すなわち個々の単語の意味からは全体の意味が計算できない表現が主に取り上げられていた。しかし日英翻訳ということを考えると、日本語で慣用表現とは言えなくても、それを英語に訳すときには単語対単語の対応以外の情報を用いて句全体で処理する必要がある表現も多くある。そこで筆者らは、日英翻訳という観点から見た日本語慣用表現として以下の4種類を扱うこととした。

1. 全体の意味が構成語の意味から計算できない表現
   例) 「油を売る」=「さぼる」=to idle away one's time
2. 機能動詞表現
   例)「電話をかける」=「電話する」=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)「ペンキを塗る」 =paint

1.は従来から言われている意味での典型的な慣用表現であり、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。2.は、体言部分が句全体の意味を担っていることが特徴である。3.は、複数の日本語単語に対して、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定まる。4.は日本語も英語も従来の定義からすると慣用表現ではないが、複数の日本語単語の組み合わせ全体に対して英語表現が対応するため、翻訳の際には句単位で処理する必要のあるものである。

慣用表現には、全体で名詞句、副詞句、動詞句、形容詞句など、様々な品詞として振舞うものがあるが、今回筆者らは、「体言(名詞 or サ変語幹)+格助詞+用言」という形式の用言句相当のものを収集の対象とした。以下、筆者らが行った慣用表現辞書の収集と、対訳付与の手順について述べる。

3. 日本語慣用表現の収集と対訳付与

3.1. 日本語慣用表現の収集

日本語慣用表現の収集に際して、まずは慣用表現を構成する語句の傾向を知るために、従来の研究[1,2]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分析を行った。その結果、体言部では、出現した語数(異なり数)の約4割が「手」などの身体の部分を表す名詞であり、約3割が「攻撃(する)」の「攻撃」などの動作

や行為を表す動作性名詞であった。一方用言部では、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調査したところ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もの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筆者らは、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コーパスから抽出し、それらの中から慣用表現の意味分布図[8]などを参考にして人間が慣用表現をピックアップするという作業を行った。これにより約8,000 表現の日本語慣用表現を得た。

さらにこれに加えて、「連絡する」の意味で「連絡をとる」と言うなどの、名詞が句全体の主要な意味を担い動詞は補助的に働く機能動詞表現[2]の収集も行った。機能動詞表現は動作性名詞と機能動詞からなっているので、収集の準備として、日英機械翻訳用辞書にあるサ変語幹約 17,500 語と和語動詞の連用形約 10,500 語の合計約 28,000 語を動作性名詞のリストとし、機能動詞としては、従来文献[2]を参考にして和語動詞約 100 語をリストアップした。これらの組み合わせを作業者に提示し、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抽出した。この方法で機能動詞表現を約 12,000 表現収集した。ただし、「サ変語幹+をする」は、サ変動詞としての枠組みで処理することができるため、今回は収集の対象とはしなかった。

3.2. 英訳語の付与

一般に慣用表現は全体で訳語が定まる特殊な表現であり、基本的には人間が訳語を付与するしかない。しかし、収集した機能動詞表現中の約 7,000 表現については日本語の言い替えを利用した対訳付与が可能であった。その他の慣用表現に関しては約 10,000 表現に人間が対訳を付与した。またこれに加えて、機械翻訳用英日辞書に着目して特定の条件に当てはまる単語を抽出することで、前節の日本語慣用表現収集手順では得られなかった約3,000 表現の対訳付き慣用表現データを作成した。

3.2.1. 言い替えを利用した対訳付与

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基本的な意味を担っており、動詞はサ変語尾「する」などと同様に補助的に機能している。従って機能動詞表現の訳語は、構成素の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、機能動詞表現を言い替えた動詞(サ変動詞または和語動詞)の英訳語を既存の機械翻訳用辞書から抽出し、その訳語が機能動詞表現の訳語として妥当か否かを人間が判断した。この場合、機械翻訳に使用する種々の情報も基本的にそのまま使用できる。ただし、日本語格フレームの助詞の値には変化がある場合があるので、注意が必要である。

3.2.2. 英日辞書を用いた日英対訳の収集

2 章で、日本語も英語も従来の定義からすると慣用表現とならない通常の表現であっても、日英の対応が一対一ではない表現に関しては慣用表現として扱うことを述べた。日英翻訳では特に、日本語が体言+助詞+用言という形式をしていて、英訳語が一語の動詞に相当する表現が多く出現する。この種の表現にはこれまでに述べた機能動詞表現のものも多いが、例えば「ペンキを塗る」="paint"のように日本語での体言が動作性名詞ではないものは、ここまで説明した手順では収集できない。そこで筆者らは、逆方向の英日辞書を活用して、以下の手順で新規の対訳付き慣用表現データを作成した。

1. 機械翻訳用英日辞書で、名詞と動詞の両方の用法がある英単語を対訳の形で取り出す。
   例)label (n.)→ラベル, label (v.)→ラベルを貼る,
   label (v.)→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、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(不採用)
3. 日本語が一語のものは対応する慣用表現を考える。

この手法で、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。

3.3. シソーラスを利用した増強

収集した約 2 万表現の日英対訳慣用表現に関して、日本語シソーラスを利用して名詞を類語表現に展開し、それぞれが慣用表現として成立するかをチェックすることで、さらに1 万表現を増強した。これにより、日本語見出し数で約3 万の日英対訳慣用表現辞書を構築した。

4. カバレッジ評価

収集した慣用表現の中で、特に体言+格助詞+動詞という形をした動詞句相当表現について、17 万文の新聞文[11]を対象としてカバレッジ評価を行った。日本語の基本動詞 100 語を選定してこれらの動詞を含む動詞句相当表現をコーパスから抽出し、慣用表現として扱うべきものの数と、それを構築した慣用表現辞書がどれくらいカバーしているかを以下の手順で調査した。

始めに、対象コーパスの形態素解析を行い、評価対象の動詞として、3 つの辞典[13,14,15]でいずれも基本動詞として扱われているものの中で、対象コーパス中での頻度の高い 100 語(「する」を除く)を選定した。これらの基本動詞は一般に多義であり、翻訳の際には訳語選択が大きな課題となる。

次に、これらの動詞を含む体言+格助詞+動詞という形をした動詞句をコーパスから抽出した。今回は、体言としては名詞またはサ変語幹、格助詞としては「が」「に」「を」に限った。また、動詞句相当表現の認定に際しては構文解析は行わず、形態素列の中から、「体言と格助詞が隣接する」「格助詞と動詞の間には 0〜4 個の用言以外の形態素の挿入を許す」というパターンに当てはまるものを抽出して動詞句相当表現とみなした。

このようにして抽出した動詞句相当表現から、慣用表現として扱うべきものと、その中で筆者らの約 3 万表現の慣用表現辞書に登録済みのものをカウントし、カバー率を計算した。表 1 に、100 動詞全体と、頻度の高い上位 12 動詞に対する本評価の結果を示す。構築した慣用表現辞書はコーパス中に出現した慣用表現の 8〜9 割をカバーしており、辞書の有効性が確認できた。特に、異なりでも延べでも上位 12 動詞でのカバー率が全体よりも高くなっており、効率的な収集を行うことができたことを示している。

表 1　カバレッジ評価結果

異なり	コーパス中の慣用表現数	その中で既登録の慣用表現数	カバー率(%)
100 動詞	1,465	1,134	77.4
12 動詞	831	717	86.3

延べ	コーパス中の慣用表現数	その中で既登録の慣用表現数	カバー率(%)
100 動詞	7,191	6,030	83.9
12 動詞	5,344	4,958	92.8

5. 英文作成支援ツールへの実装

5.1. FEP 型英文作成支援ツール「英文名文メイキング」

筆者らが開発した英文作成支援ツール「英文名文メイキング」は、日本語の単語、句、文を英語に変換する英文作成 FEP「英作ペン」、CD-ROM 辞書を検索する「書籍レンズ」、日英対訳例文の利用を支援する「例文バインダー」の 3 つのツールからなる。「英作ペン」は二重のフロントエンド機構により、任意のアプリケーション、かな漢字変換 FEP と組み合わせて使用することができる。ユーザは日本語を入力して変換→選択→確定というサイクルを繰り返すことで目標とする英文を対話的に作成する。ユーザが変換したい日本語を入力すると自動的に日英変換ウィンドウが立ち上がり、入力がリアルタイムに形態素解析されて自立語が自動抽出され、英語に変換される。図 1 に、「英文名文メイキング」の変換ウィンドウの画面例と、ツール全体の概要を示す。ユーザはここで、訳語や文型の変更、各単語の活用情報などを利用することができる。さらにこの変換ウィンドウから、慣用表現辞書を含むシステム辞書、例文、CD-ROM 辞書など、英文作成のための種々の関連情報に自在にアクセスすることができる。この後、日英構文変換を指示すると、語順変更、冠詞挿入などの処理が行われて、全体が英文に変換される。

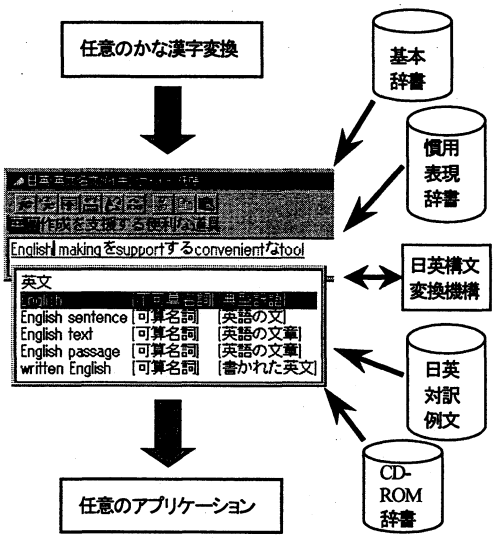

図 1　「英文名文メイキング」の概要

5.2. 慣用表現の訳語選択インタフェース

慣用表現には一般に、慣用表現としての意味と文字どおりの意味の両者が存在する。例えば「手を切る」という表現には、「関係を絶つ」という慣用表現としての意味と、「手を怪我する」意味が存在する。従って翻訳の際には、構築した慣用表現辞書に従って慣用表現として翻訳することができるだけではなく、両者の訳語を選択できるようになっていることが望ましい。筆者らの開発した英文作成支援ツールでは、単語等に対する訳語候補選択と同様に、慣用表現をワンタッチで分割/再変換することが可能なインタフェースを実現している。例えば「油を売る」という入力に対しては、本ツールはまず全体を慣用表現として認定し、図2の上に示すように変換する。ここで、慣用表現に対する各種の訳語は、単語に対するときと同様に任意に選択することができる。

これに加えてここでは、「油/を/売る」という"/"を含んだ日本語表示部分を選択することで、慣用表現ではない文字通りの意味も選択することができるようになっている。マウス操作やカーソルキーで「油/を/売る」を選択すると、図2の下のように「油」と「売る」が個別に変換される。また逆に、この状態から「油+を+売る」という"+"を含んだ日本語部分を選択することで、容易に慣用表現としての訳語を再度選択することができる。

図2 慣用表現の日英変換例

6. おわりに

構築した日本語見出し数で約3万の日英対訳慣用表現辞書について、収集・辞書構築方法の概要と新聞文を対象としてカバレッジ調査を行った結果を報告した。日本語の基本動詞100語を含むコーパスに出現する慣用表現に関して、構築した慣用表現辞書は8割強をカバーしており、辞書の有効性が確認できた。また、英文作成支援ツールで実現している、慣用表現としての意味と文字通りの意味を簡単に選択するためのインタフェースについても説明した。今後は、語彙をさらに拡充するとともに、対訳コーパスを利用した評価も行って、機械翻訳システムの品質向上を図っていく計画である。

参考文献

[1] 宮地: "慣用句の意味と用法", 明治書院, 1982.

[2] 村木: "日本語の機能動詞表現をめぐって", 国立国語研究所報告65, 1980.

[3] 奥: "日本語慣用表現の分析と日英翻訳への適用", 情報処理学会自然言語処理研究会, 87-NL-62-2, 1987.

[4] 首藤 他: "日本語の慣用的表現について", 情報処理学会自然言語処理研究会, 87-NL-66-1, 1987.

[5] 新納 他: "語義の特異性を利用した慣用表現の自動抽出", 情報処理学会論文誌, Vol.36 No.8, 1995.

[6] 田中: "大規模コーパスを用いた共起関係の抽出", 言語処理学会第3回年次大会, 1997.

[7] 桑畑 他: "IPAL 辞書統合による多義性解消のためのコロケーションの分析", 言語処理学会第3回年次大会, 1997.

[8] 亀井 他: "日本語の用言相当慣用表現の意味空間における分布図", 言語処理学会第3回年次大会, 1997.

[9] 田村 他: "日英機械翻訳のための大規模慣用表現辞書の構築", 言語処理学会第2回年次大会, 1996.

[10] 亀井 他: "人間の言語知識と機械支援を融合させる機械翻訳インタフェース Source and Target language Mixed Stage", 言語処理学会第3回年次大会, 1997.

[11] 日本経済新聞社: "日経全文記事データベース日本経済新聞CD-ROM 1994年版"

[12] 土井 他: "FEP 型英文作成支援ツール ―訳語選択のユーザインタフェースと辞書記述―", 情報処理学会第51回全国大会, 5H-1, 1995.

[13] 金田一 他: "新明解国語辞典", 三省堂, 1972.

[14] 森田: "基礎日本語辞典", 角川書店, 1989.

[15] 小泉 他: "日本語基本動詞用法辞典", 大修館書店, 1989.