用言句相当慣用表現辞書のカバレッジ評価

7L－11

田村 真子　亀井 真一郎
ＮＥＣ 情報メディア研究所

1. はじめに

複数の特定の単語が組み合わさって句全体で特別な意味を生じる表現(以下慣用表現と呼ぶ)を正しく翻訳することは、機械翻訳システムにおける訳語選択の曖昧性の問題を解決する課題の一つである。そこで我々は「さほる」の意味での「油を売る」のように従来研究で慣用句と呼ばれている表現の他に、「電話をかける」に対応する「make a phone call」のように翻訳の際に訳語選択が問題となるようなコロケーションも対象として日本語見出し語数で約20,000の日英機械翻訳用慣用表現辞書を開発した[1]。

今回、一般に多義のため翻訳の際に訳語選択が問題となるような13動詞を選び、新聞記事約170,000文(約40Mbyte)[2]に出現するそれらの動詞を含む慣用表現を、我々の辞書の見出しがどれくらいカバーしているかの評価を行なったので報告する。また、個々の動詞句における慣用表現の頻度の分布も調査したのでその結果も報告する。

2. カバレッジ評価

2.1. 対象動詞の設定

まず、３つの市販辞典[3, 4, 5]を選定し、それら全ての見出しとなっている動詞をピックアップした。今回はそれらのうち新聞記事約170,000文で出現頻度が300以上である13の動詞を評価対象として選んだ。表１は13の動詞の一覧である。

2.2. 動詞句データの収集

今回、以下のような手順で動詞句データを作成した。

1. 以下のような単語列をプログラムによりコーパスから抽出する。

   (A) (B) (C) (D) (E) (F) 動詞 (G)

   A～Gはそれぞれ単語であり、以下の条件を満たすものとする。

   (a) A～Fの間に「体言, 格助詞」の連続を含む。
   (b) 「体言, 格助詞」と動詞の間に別の用言を含まない。

   なお今回、体言としては名詞あるいはサ変語幹を取り、格助詞は「が」、「を」、「に」に限った。
2. 1.の結果ファイルから前節で選択した動詞を含む単語列を抽出し、動詞ごとにファイルを作成する。

ある動詞と係り受け関係にある単語列をその動詞の前方に検索してゆくのは解析の曖昧性などの問題により困難である。そこで今回我々は上記1.のように、動詞に係る単語列を含むと思われる範囲を予め設定しておき、その範囲内でデータ分析を行なうことにした。今回は動詞の前方の単語を６単語とした。

また、従来なされているように[6]［体言, 格助詞, 動詞］の共起データを、それが出現する文から切り離して取り出すことをせずに上記1.のように［体言, 格助詞, 動詞］の並びを含む単語列という形式で動詞句データとしたのは、評価の際に原文に戻って分析をしやすくするためである。また、上記2.において、AからFの単語列の間に出現する［体言, 格助詞］を全て拾ったが、この点も従来の方法[6]と異なる。

2.3. 評価

評価は以下の手順で行なった。

1. 前節で得られた13個の動詞句ファイル個々において、当該動詞を軸として「体言, 格助詞」の組をコーパス中の出現頻度ごとにクラスタリングする。
2. 「体言, 格助詞」の組の中で我々の慣用表現辞書の見出しとなっているものにマークをつける。
3. 我々の慣用表現辞書の見出しとなっていないものの中で慣用表現であるものにマークをつける。
4. 2.と3.の結果を元にカバレッジの割合を算出する。

表２は上記1.の結果ファイルの例であり、動詞句を構成する体言を頻度順に並べたものである。表中の「＊」は慣用表現辞書の見出しであることを、また「！」は慣用表現であるが辞書見出しにはないことを表している。

表３は、上記4.の結果、すなわち今回の評価結果である。表中のＡ欄からＤ欄はそれぞれ、A: コーパス中で当該動詞を含む慣用表現の全出現頻度、B:A欄のうち我々の辞書に登録されている表現の全出現頻度、C: コーパス中の慣用表現の全異なり数、D:C欄のうち我々の辞書に登録されている表現の全異なり数を表している。また表中Ｂ欄とＤ欄の()内の数字はそれぞれ、Ｂ欄のＡ欄に対する割合、Ｄ欄のＣ欄に対する割合それぞれ表す。例えば、「～が出る」の形式の慣用表現はコーパス中に538回出現しておりそのうち我々の辞書の見出しであるものは378回(70.3％)出現している。また、異なり数で見るとコーパス中には60件の慣用表現が出現し、そのうち54件(90.0％)が辞書に登録されている。

表３から、新聞文において13の動詞を含む慣用表現について我々の辞書見出しは(1)コーパス中の出現頻度から

Evaluation of the Dictionary of Japanese-English Idiomatic Expressions
TAMURA Shinko, KAMEI Shin-ichiro
NEC Information Technology Research Laboratories

見て約8.8割をカバーしていること、(2)異なり数で見ると約8.2割をカバーしていることが分かる。

3. 考察

以下、今回の評価の結果について考察する。

我々は慣用表現対訳の収集にあたり、それが全くの手作業となることを避けるために慣用表現の構成要素の言語的な特徴を用いたり、英日辞書を利用するなどの工夫をした[1]が、今回のような頻度データは用いず、人の直感に拠るところが大であった。また、従来研究で実際のコーパス中の慣用表現の頻度の分布を調査した報告はなかった。

そこで今回、13件の高頻度動詞において［体言,格助詞,動詞］の並びの頻度データをとり、それぞれの動詞において頻度の高いものから上位10表現を調べたところ、平均して7表現が慣用表現であることが分かった。このことと今回対象とした13件の動詞がコーパス全体で高頻度で出現することを考え合わせると、コーパス中の句表現のうち慣用表現が占める割合は決して小さくはないことが推測できる。従って今回の調査から慣用表現を適切に翻訳することは機械翻訳の質の向上に大きく寄与することが予測できる。

一方、中頻度ないし低頻度の表現の中でも慣用表現が一定の割合で出現していることも確認した。例えば、「身を入れる」などの身体の一部を表す名詞を含む表現や、「アルコールが入る」などの日常生活に密着した表現は、今回の評価データでは頻度1であった。頻度に重点を置いた収集法では見過ごされがちなこのような表現が既に辞書化されていたのは頻度に拠らない収集作業の結果である。今回は新聞文を対象としたので頻度が低かったものの、このような表現は日常的な文章にはより高頻度で現れそうな表現であり、今後も我々はこのような表現の辞書化を進める。

以上の考察から、慣用表現のカバレッジを向上するためには、従来なされてきたような人の直感を手掛かりとして収集する方法の他に、今回のような頻度データを用いて高頻度の表現を収集する方法を積極的に取り入れる必要があることが言える。

また、本来慣用表現とすべきであるのに我々の辞書に見出しがなかった表現には以下のような表現が多くあった。

1. 体言が辞書見出しの体言の同義語や類義語である
   (例) 視点に立つ (辞書に有)
   v.s 視野に立つ (辞書に無)
2. 表現全体が辞書見出しの同義語や類義語である
   (例) 視野に入れる (辞書に有)
   v.s. 考慮に入れる (辞書に無)
3. 体言部分が辞書見出しの体言の下位分類である
   (例) 電話を入れる (辞書に有)
   v.s. 電話連絡を入れる (辞書に無)

今後は、上記のような同義語や類義語の情報を手掛かりとして、既に収集した表現を元に慣用表現を追加収集する必要がある。

4. おわりに

新聞文を対象として我々の開発した慣用表現辞書のカバレッジ評価を行なった。その結果、コーパスに高頻度で現れる13件の動詞を含む慣用表現について、我々の慣用表現辞書は、コーパス中の出現頻度から見ても慣用表現の異なり数で見ても8割を越えるカバレッジであった。また、今回の評価分析の結果、(1)コーパス中に高頻度で出現する句単位の共起データに慣用表現が多く含まれていること、(2)一方で低頻度のデータにも収集すべき慣用表現も多くあることが分かり、慣用表現のカバレッジを向上させるためには、従来の人の直感に頼っていた収集法に加えて、今回のような頻度情報を用いて高頻度の表現を収集する方法も積極的に取り入れることが必要であるという知見を得た。今後はこの知見の他、同義語や類義語の利用も考慮して辞書開発を進める予定である。

表1:評価対象の動詞

与える	入れる	受ける	かかる	得る	置く	立つ	
つく	出る	乗る	入る	結ぶ	持つ		

表2:「〜に立つ」の結果ファイル(一部省略)

*役(25)	*教壇(7)	…	!視野(3)
*立場(22)	!上(6)	!反省(4)	…
*先頭(15)	!頂点(6)	*間(3)	*ステージ(1)
*トップ(13)	売上(6)	*窮地(3)	*ピンチ(1)
*岐路(12)	…	*首位(3)	*瀬戸際(1)
!挨拶(11)	!展望(5)	*転機(3)	…
*視点(10)	面(5)	!最前線(3)	住民(1)
*舞台(10)	*先(4)	…	!謝辞(1)

表3:評価結果

	A	B		C	D	
…	…	…	…	…	…	…
(を)得る	324	322	(99.4)	37	35	(94.6)
(に)入れる	125	52	(41.6)	18	12	(66.7)
(を)入れる	458	446	(97.4)	35	26	(74.3)
(が)出る	538	378	(70.3)	60	54	(90.0)
(に)出る	90	71	(78.9)	23	20	(87.0)
…	…	…	…	…	…	…
合計	5,124	4,527	(88.3)	722	591	(81.9)

A: 当該動詞を含む慣用表現の全出現頻度
B: 辞書見出しの慣用表現の全出現頻度
C: 当該動詞を含む慣用表現の全異なり数
D: 辞書見出しの慣用表現の全異なり数
() 内は慣用表現辞書見出しの占める割合

参考文献

[1] 田村,亀井:“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.

[2] 日経全文記事データベース日本経済新聞CD-ROM版1994年版.

[3] 金田一京助他:“新明解国語辞典”,三省堂,1972.

[4] 森田良行:“基礎日本語辞典”,角川書店,1989.

[5] 小泉保他:“日本語基本動詞用法辞典”,大修館書店,1989.

[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”,情報処理学会論文誌,Vol.36,No8,1995.