共起対象名詞を共有する動詞間の関係

－分類と利用－

斎藤佳美　　野上宏康

（株）東芝　研究開発センター

1.はじめに

これまで、様々なかたちでコーパスデータの利用の必要性が認識されており、その中でコーパスデータから得られる語彙知識のひとつとして共起に関する知識の利用が提案されている。例えば、語彙の意味的なクラスタリング［1］［2］、文書検索のための類義語知識の獲得［3］、多義動詞の解釈知識の獲得［4］、定型表現の抽出［5］などの研究がある。

コーパスデータ、共起データの利用研究においては、ベースとなるデータの質・量によって獲得できる知識の質・量が大きく左右されざるを得ない。そのためか、語彙知識獲得の対象が高頻度で安定して出現する語彙に限定されることが多かった。だが実際には、中頻度・低頻度の語彙の知識獲得のためにコーパスデータを利用したい場面は少なくない。最近のように大規模なコーパスデータ、共起データの利用が可能になれば、さらにその期待は高まる。

ただしその際には、中頻度、低頻度の語彙について、共起データ中での“位置”を改めて確認する必要があるように思われる。中でも、これまで共起データの利用は名詞の分類を中心に検討されてきており、動詞の分類についての報告は少ない。

そこで我々は、共起データの利用において、中頻度以下の動詞に関し、どのような語彙知識、意味分類知識を獲得できるかの検討を試みた。

共起データとしてはＥＤＲ共起辞書［6］を用い、共起現象としての特徴が顕著で扱いやすい、“を”格に名詞が出現する場合について分析を行った。

2.意味分類の操作

意味分類作業の内容には、主に次の２つの操作が含まれていると思われる。

- 上位（または下位）の語彙を見つける。
- 意味の類似性を持つ語彙を見つける。（これを以下、同類の語彙と呼ぶこととする。）

これまで、共起データ利用の研究においては、このうちの同類の語彙の発見が主な検討課題であり、今回の我々の検討もこの観点を中心に行った。

同類の語彙としては、少なくとも同義語、反義語、類義語などが想定される。「建設する」という単語の例でいえば、図１に挙げたような語彙である。

図１　同類の語の例

「建設する」の場合

反義語：破壊する、解体する

類義語：建築する、建てる

上位語：つくる

下位語：建造する

3.共起データの統計的分析

意味分類獲得の検討を行う前に、その準備として、まず共起データの統計的な特徴を分析した。

ＥＤＲ共起辞書中に登録されている「を」共起データは、異なりで約 85000 件、のべ数では約 125000 件であった。（斎藤調べ）このうち出現回数が 5 回以上の共起表現が約 5%、出現回数１回の共起表現が全体の約半数を占めている.

ここで、各名詞／動詞が共起する単語の数に注目し、調査した結果を示したのが図２である.（ただし、名詞と共起する動詞の中でサ変動詞の「する」、動詞と共起する名詞の中で「これ」「それ」「こと」「もの」は除いた。また、１種類の動詞としか共起しない名詞、１種類の名詞としか共起しない動詞も除いてある。）横軸に各語彙毎の共起対象単語の数を、縦軸に該当する単語（名詞／動詞）の度数を表示している。

また共起対象単語数の１位～１０位までの動詞と名詞、各１０単語、対象単語数の平均付近にあたる例（動詞では単語数 13、名詞では単語数 7）各１０単語、対象単語数２の動詞と名詞の例各１０単語を図３に挙げる。

図２－ａ）共起対象単語数の分布　その１

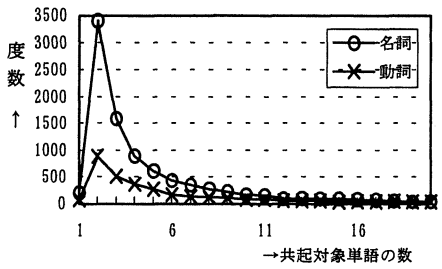

図２－ｂ）共起対象単語数の分布　その２

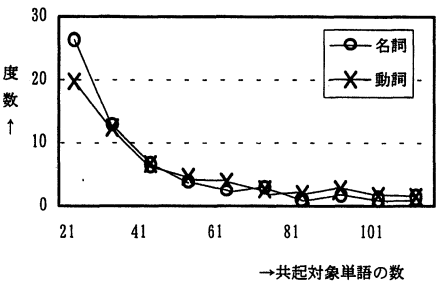

図３共起対象単語数の分布 ---単語例---

動詞

1位～10位：持つ、見る、いう、使う、求める、
受ける、行う、出す、示す、作る

対象単語数 13：印刷する、つづる、目撃する、
保障する、閉める、分離する、
連想する、増大する、 満載する....

対象単語数 2：睨む、猶予する、露呈する、
連載する、歴訪する、編み出す、
郵送する、立案する、紛らす、
払い戻す、吐露する、.....

名詞

1位～10位：情報、データ、問題、関係、人、
機能、システム、手、部分、時間

対象単語数 7：和平、輪郭、離婚、問い、役員、
募金、文献、付近、発言力、農場、.....

対象単語数 2：賄賂、録音、連絡帳、列車番号、
冷たさ、臨時国会、良否、理事長、
流域、裏手、.....

ここで、共起対象単語数と語彙の意味分類に関して次のような仮定を置き、共起対象単語数による上位、中位、下位の区分を設定することとする。

- 度数が上位の単語は、概念体系上も上位に位置する。
- 度数が中位～下位の単語は、概念体系上も中位・下位に相当する。

すなわち、仮に、およそ1：10：100という比率で動詞を3層に分けるとすると、対象単語数201以上が上位、35～200が中位、34以下が下位という区分になり、語彙数はそれぞれ約40単語、約360単語、約3500単語となる。中頻度以下の動詞は、主にこの区分における下位の範囲に含まれることになる。

4.共起対象名詞を共有する動詞の分析

前述の統計的特徴を踏まえ、下記の手順で共起データから共起対象名詞を2つ以上共有する動詞の抽出を試みた。

（1）ＥＤＲ共起辞書から、目的の動詞と共起している名詞を取り出す。
（2）これら名詞と共起する動詞を見つける。
（3）はじめの動詞と共起していた名詞群のうちの複数の名詞と共起している動詞を取り出す。

この方法で、先に述べた定義によれば下位となる、（図3に挙げた）対象単語数13の動詞（合計48単語）について共起対象名詞を共有する動詞を集めたところ、平均で約52単語が抽出された。抽出例を図4(a)に示す。

例えば「印刷する」では、対象名詞の数が同程度（34以下）の動詞は集められた動詞の約2割にとどまり、意味が類似すると思われるのは「刷る」の1語のみであった。また、相対的に上位に当たる動詞（対象単語数34～

図4 共起対象名詞を共有する動詞

(a)＜対象単語数34以下の動詞：約3500単語＞
「印刷する」：共起対象単語数 13　(3)の動詞数 51
対象単語数34以下　11単語(一致名詞数順)
入力する、読み取る、送受信する、書き込む、刷る、出力する、刻む、検索する、添える、点検する、組み立てる
対象単語数で35～200　29単語 (一致名詞数順)
表示する、知る、書く、読む、通す、出す、指定する、使う、発表する、選ぶ、置く、調べる.....

「つづる」：共起対象単語数 13　(3)の動詞数 51
対象単語数34以下　12単語（一致名詞数順）
詠む、入力する、振り返る、しゃべる、区切る、共有する、分類する、込める、ぶつける、歌う、体験する、募る
対象単語数で35～200　26単語(一致名詞数順)
通す、表現する、読む、語る、認識する、聞く、伝える、通じる、送る、書く、使う、覚える、.....

「目撃する」：共起対象単語数 13　(3)の動詞数 37
対象単語数34以下　9単語（一致名詞数順）
報道する、追跡する、想像する、映す、撮る、あばく、思い浮かべる、浮かべる、再現する
対象単語数で35～200　20単語(一致名詞数順)
伝える、描く、表現する、通す、追う、知る、思い出す、起こす、報告する、ながめる.....

(b)＜対象単語数35～200の動詞：約360単語＞
「覚える」：共起対象単語数 97　(3)の動詞数 358
対象単語数で35～200　192単語(一致名詞数順)
感じる、教える、呼ぶ、学ぶ、抱く、忘れる、伝える、書く、調べる、.....
対象単語数で201以上のもの　33単語(一致名詞数順)
持つ、使う、受ける、示す、続ける、求める、用いる、考える、利用する、作る、.....

200）は集まった動詞の約半数を占め、上位の意味を持つと思われる単語は含まれていない。この傾向は他の動詞でも同様であった。

比較のため同様の手順で中位の動詞に対して行った例を図４（b）に示す。中位の動詞では、「覚える」に対して「忘れる」「感じる」などの反義語、類義語に当たる語彙の多くを集められた動詞の中に発見することができた。下位の動詞で同類の語彙の知識があまり得られない理由は、そもそも同類に認定できる語彙が少ないのか、あるいはその他の問題によるのか、この点は今後の検討課題である。

一方、下位の動詞に対しても、同類の語彙に代わる情報は獲得可能であると言える。すなわち図４に示した動詞間の関係を見ると、“一連の流れ”を持った動詞のまとまりが少なからず存在している。例えば「印刷する、表示する、入力する、送受信する、書き込む、読み取る」「つづる、詠む、体験する、振り返る、しゃべる」「目撃する、報道する、追跡する」などである。

これらの動詞間には、ある出来事の様々な側面の関係や時間的な前後関係などを見ることができ、個々の２動詞間での意味の類似性は薄いが、組み合わせとして考えると使用場面を共有するものと考えられる。「印刷する、入力する、表示する、書き込む、読み取る」の例では、これらが共有した共起名詞は「文字」「情報」であったが、「印刷する」を除いた「入力する、表示する、書き込む、読み取る」のうちの３動詞が共有する共起名詞は「データ」「画像」で、「印刷する」ものとしても適当な単語となる。

共起対象単語数 13 の動詞の場合、抽出された動詞のうち、このような“一連の流れ”を持つと思われる動詞の割合は約３割であった。

5.おわりに

動詞の中には、語彙の意味から類義語、反義語などを設定することが難しい単語も少なくない。大規模な共起辞書（共起データ）は動詞の語義に関する知識を得るのに有効な情報となる。

もちろん先に上げた手順だけで必要十分な知識が得られるわけではない。その理由のひとつは手順の単純さにあるが、共起データの偏りや量的な熟度の不足により十分な分解能が得られない点も考えられる。このような検証のためにも今後共起データとして中頻度以下の語彙の使用用例の蓄積が望まれる。

今後とも、語彙知識獲得の基礎となる共起データについて質・量の数値化、視覚化、共起辞書間での比較評価などが検討課題であると考えている。

参考文献

[1] 藤原祥隆：“共起パタン分布に基づく単語間類似度を用いた動詞・名詞のクラスタリング法”,第２回人工知能学会予稿集,8-3,1988

[2] 井ノ上直巳、森元つよし：“クラスタリング手法と既存シソーラスとの組み合わせ手法”,情報処理学会第 42 回全国大会,7E-6,1991

[3] 下村秀樹、福島俊一：“共起類似性に基づく同義語の抽出”,情報処理学会第 47 回全国大会,1M-10,1993

[4] 平岡冠二、松本裕治：“共起情報を用いた多義動詞の類別と名詞のクラスタリング”,言語処理学会第１回年次大会,1995

[5] 新納浩幸、井佐原均:： 片方向の共起性による述語型定型表現の抽出”,言語処理学会論文誌,Vol.2,No.3,pp723-85,1995

[6] ＥＤＲ：“日本語共起辞書”,（株）日本電子化辞書研究所,1995