

表層格と動詞の関係に基づく動詞の自動分類

足立 顕 牧野 武則

東邦大学 大学院 理学研究科

〒274-8510 千葉県船橋市三山町2丁目2番1号

E-mail: akira@mk.is.sci.toho-u.ac.jp

makino@is.sci.toho-u.ac.jp

あらまし

本研究は、表層から得られた情報だけを用いて、動詞を分類し、分類結果を基にして格と格の関係を見つけ出すことを目的とした。そのために、自由格と動詞との関係を表層で表しているもののうち格助詞相当語句が限定した意味・概念を持ち動詞と関係を持っている点に着目した。格助詞相当の働きをする語は新聞記事コーパスから抽出し、その語と動詞の共起頻度を用いてクラスタ分析を行った。

本稿では、自由格を用いた動詞の分類の有効性と、格と格との間の関係を分類結果よりモデル化したものについて報告する。

Classification of Japanese Verbs Based on Surface Case.

Akira Adachi and Takenori Makino

Graduate School of Information Science, Toho University,

Chiba 274-8510, Japan

Abstract

In this paper, We report effectiveness of classification verbs based on surface case makers including verb phrases regard as case marker. The surface case marker extracted automatically from news paper corpus without subjectivity, and it relate to verbs with limited semantic features.

The classification means a relationship deep case of verbs with disambiguity. We discuss the relationship among semantic case marker and verbs classification obtained by method of clustering algorithm.

1 はじめに

高度な言語処理を行うためには、より充実した辞書の整備を必要とすることはいうまでもない。そのため、EDRをはじめとして電子化辞書の開発が人手により行われてきた。[1]

辞書の開発は非常に多くの労力と時間を必要とすると同時に、開発者の主観的な影響を受け記述の一貫性を維持することが難しい作業である。[2] そのため、大規模コーパスを用いて統計的に辞書を生成しようとする研究が行われている。

これまでの研究の中心は、コーパスを解析することで得られる動詞に対する格助詞のパターン、即ち必須格の情報を統計的に処理することで動詞の概念分類を行っている [4][5]。

しかし、文脈より推測できる格は、格が省略されることが多くコーパスから正確にパターンとして動詞との関係を得ることは難しい。また、パターンとして得られたとしても、表層で関係を表している格助詞は多様な用法をもっているため分類結果の安定度について疑問が残る。

本研究では、文脈上任意に用いられる格助詞相当の働きをする助詞（格助詞相当語句）が動詞と安定した概念関係を持つという点に着目し、クラスタ分析を用いて動詞の概念分類を行った。

本稿では、格助詞相当語句をコーパスより自動的に抽出する方法およびその情報を利用し動詞を概念分類した結果について報告する。

なお、本研究では、コーパスとして日本経済新聞社記事データベース1998年度版を用い、日本語形態素解析システムには「茶筌」Version2.02を利用した [6]。また、実験に必要な語彙（格助詞相当語句など）については、本稿で述べる手法によりコーパスより獲得し茶筌の辞書に変換して実験を行った。

2 格助詞相当語句

2.1 格助詞相当語句の性質

「格助詞」+「動詞（連用形）」もしくは「格助詞」+「動詞（連用形）」+「接続助詞」に

より構成されている語は、複文における並列節の総記の並列を作る用法と同じである。これらの並列節には、連用並列とテ形並列があり、並列表現に加え、原因、理由、手段、付帯状況などを表す副詞節を表す用法を持つ。

連用並列とテ形並列の違いは、連用並列がテ形並列より文語的といえ、また、テ形並列の方が連用並列よりも時間的前後関係が明確なという基本的性質を持つ。[3]

● 連用並列

- － 雨が降り、雷がなった
- － 雷がなり、雨が降った

● テ形並列

- － 雨が降って、雷がなった
- － 雷がなって、雨が降った

これら連用並列、テ形並列は、格助詞相当の機能を持ち、格助詞を用いて、次のように言い換えることができる。

- 雨で雷がなった
- 雷で雨が降った

この場合、原因・理由を意味する格助詞「で」が用いられる。

同様にして新聞コーパスより抜き出した並列節の具体例について考察する。

- 「に比べ」「に比べて」→「比較」を意味する格助詞「より」を用いて言い換えることができる
 - － 前年産に比べ最大で約一六%上がるなど高値でスタートした（前年産より最大で）
 - － 日本人は米国人に比べてユーロに鈍感だ（米国人よりユーロに）
 - － 前年同期に比べて売り上げが減少、歳暮の売り上げも落ち込んだ（前年同期より売り上げが）
 - － 売上高、経常利益とも九八年実績見込みに比べ増加企業の比率が高くなっている（実績見込みより増加企業の）

- 百貨店も前月（四・八％減）に比べて落ち込み幅が縮まった（前月より落ち込み幅が）

「比べる」そのものに「比較」を示す概念を備えているため、対象を表す格助詞「に」を伴い比較可能な概念を持つ動詞と共起する。

- 「に向け」「に向けて」→「方向・目的・終状態」を意味する格助詞「へ」を用いて言い換えることができる

- 混迷を振り切り二十一世紀に向けダッシュする年にしたい（二十一世紀へダッシュする）
- 決算期末に向けて金利上昇懸念が出ている（決算期末へ金利上昇）
- 資金不足の緩和に向けて積極的な役割を果たすよう期待を示した（緩和へ積極的な）
- 協議機関づくりに向け近く松原武久市長と直接意見交換する意向だ（機関づくりに近く）
- 復活に向けて国に強く働きかけたい（復活へ国に）

「向ける」そのものに「方向」を示す概念を備えているため、目的・方向・終状態を表す格助詞「に」を伴い、空間的、概念的・方向に対して移動可能な概念を持つ動詞と共起する。

- 「に伴い」「に伴って」→「原因・理由」を意味する格助詞「から」を用いて言い換えることができる

- 低気圧の発達に伴って北日本では冬型の気圧配置が強まる（発達から北日本では）
- 半導体の高集積化に伴って純度の高いプロセスガスの要求が高まっており（高集積化から純度の高い）
- 同工場では生産品目拡大に伴ってラインの構築を進めてきた（拡大からラインの）
- 大手スーパーが秋に新規出店に伴い大量にパート募集したことも影響した（出店から大量に）
- 初心者急増に伴い、安易な気持ちでインターネット個人輸入をする人が増えているためだ（急増から安易な）

「伴う」には、「（ある物事を）付随して生じさせる」という概念を持っている。従属節で表現された内容が理由、原因となり主節の動作に影響をおよぼしている。そのため、影響を受ける動作・変化には、「現象」「移動」「行為」などの概念を持ち、同時に起っている動作の影響を受けることができるものである。

- 「に対し」「に対して」→「対象」を意味する格助詞「に」を用いて言い換えることができる

- 金利上昇に対して時には最大三兆五千億円の資金供給を実施する（金利上昇に時には）
- 健全な取引先に対して必要な資金供給が円滑に行われない（取引先に必要な）
- この開発案件に対して条件付きながら環境評価面からは問題ないと発表（開発案件に条件付き）
- 銀行は不良債権に対して引当金を計上し、会計処理している（債権に引当金を）
- 公取委は違反事業者らに対し当該違反行為を排除するのに適当な排除措置をとるよう勧告する（事業者らに当該違反を）

「対する」には「面と向かい合うこと」という「対象に関する位置関係」を表す概念を持つ。従属節で表現された「対象」に対する動作・変化が主節で示される。

このように格助詞相当語句は、格助詞相当語句を構成する動詞の概念の一部を用いているため、格助詞単独と比べると限定した意味となり主節の動詞と関係している。

2.2 格助詞相当語句候補の抽出

格助詞相当語句は「格助詞」＋「動詞（連用形）」、「格助詞」＋「動詞（連用形）」＋「接続助詞」により構成され、その性質について述べてきた。これらの語の組み合わせは形態素解析処理を行うと表1に示すように容易に取り出すことができる。

このようにして得られた語の組み合わせには、次のようなものが含まれている可能性がある。

頻度	格助詞	動詞	接続助詞
9648	に	比べ	-
7960	を	示し	-
6856	を	受け	て
6439	に	向け	て
5794	を	進め	て
5788	を	始め	-
4553	を	まとめ	-
4332	に	比べ	て
3849	を	含め	-
3489	に	向け	て
3480	が	増え	て
3414	を	求め	-
3334	に	伴い	-
3192	に	達し	-

表 1: 格助詞相当語句候補

- 格助詞相当語句として機能するもの
- 動作の付帯状況を意味するもの
- 形態素解析辞書に未登録の慣用的に用いられる複合動詞が分割されたもの

本実験では、表層から得られた情報のみを利用して動詞の分類を行うことを目的としており、これらの語が含まれていることを認識した上で、格助詞相当語句として抽出した語の生起頻度を利用したデータの間引き以外の処理は行わず、動詞の分類を行う。

3 分類実験

2.2 より得られた、格助詞相当語句（候補）のうち、生起頻度 50 回以上の語彙（342 語）を格助詞相当語句として動詞の分類実験に採用した。

3.1 共起関係抽出

動詞と格助詞相当語句との共起関係を抽出には図 1 に示すように形態素解析と簡易構文解析を用いて行った。

簡易構文解析は、文全体を構文解析するのではなく、係り受け関係の区切りとなる条件を設定し、その内部で表記されている格助詞および

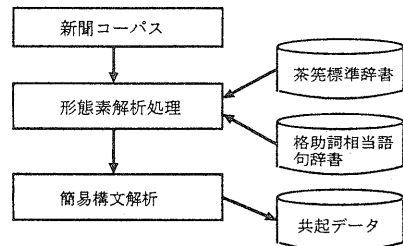


図 1: 共起関係抽出の流れ

格助詞相当語句は直後の動詞と関係を持つとした。

係り受け関係の区切りとなる条件は、非交差性（係り受けは交差しない）、唯一性（後方にあるいずれかの語にかかる）[7]を前提として、係り受けの飛躍を起こす可能性のある係り助詞および名詞の係り先となる動詞を区切りとした。

この条件を用いて構文解析を行うと図 2 のような関係を得ることができる。



図 2: 簡易構文解析結果

以上のルールを適用し新聞コーパスを解析すると表 2 に示すような結果を得ることができる。「頻度」は、動詞の格助詞との共起頻度の総数、「格助詞および格助詞相当語句」の表記の直前の数字は、各格助詞と動詞との共起頻度を意味している。

3.2 クラスタ分析

格助詞相当語句が独立して存在していると仮定し、動詞ごとに格助詞相当語句を各要素とするベクトル v_i を生成した。 v_i と格助詞相当語句 (c_k) の共起頻度 (c_{ik}) を各要素の値として採用した。

$$v_i = (c_{i1}, c_{i2}, \dots, c_{in})$$

類似度は、次の式で示すように動詞 v_i, v_j べ

動詞	頻度	格助詞および格助詞相当語句
出発	446	135, から 68, として 4, とともに 38, に向け 26, に向けて ...
前進	296	3, として 2, とともに 47, に向け 32, に向けて ...
提出	450	10, として 77, に 6, について 11, をまとめ ...
提示	132	88, が 3, にとって 6, をまとめ 1, を見直し 1, を受けて ...
提訴	636	37, として 23, を求め 48, を求めて 5, を相手取り ...

表 2: 共起関係抽出結果

クトルを正規化した上で内積値を計算し大きな値（もっとも 1 に近い値）を取るものをもっとも類似度が高い動詞の組み合わせと判断する。

$$e_i = \frac{v_i}{|v_i|} = \frac{1}{\sqrt{\sum_{k=1}^n c_{ik}^2}} (c_{i1}, c_{i2}, \dots, c_{in})$$

$$sim = \max_{i,j=1, i < j}^n (e_i \cdot e_j)$$

類似度の高い動詞の組 (v_i, v_j) からクラスタを生成する際は、クラスタを構成する動詞のベクトルから重心の計算を行い、クラスタのベクトルとした。

$$v_i = Cluster(v_i, v_j) = \frac{v_i + v_j}{2}$$

次に、クラスタ (v_i) と動詞 (v_k) により新しいクラスタを生成する場合には次の式を用いた。

$$v_k = Cluster(v_k, v_i) = \frac{v_i + v_j + v_k}{3}$$

クラスタ間の重心ベクトルについても、それぞれを構成する動詞のベクトルの和を計算した上で、そのベクトルの個数で割るという処理を行った。

このような計算を行うことで、図 3 に示すような結果を得ることができる。図の中の数値は、内積の値を意味している。

クラスタ分析を行うことで類似度の高い動詞を階層に並べることができる。一方で概念のまとまりとしてのクラスタを決定するためには、どの程度の類似度を持つクラスタに分割すれば良いかという問題がある。

そのため、次項で述べる手法によりクラスタに分割する際の閾値を求めた。

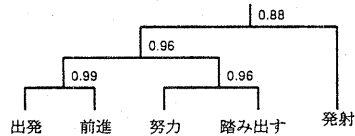


図 3: クラスタ解析例

3.3 クラスタの閾値

格助詞相当語句と動詞との共起関係が単位期間および共起頻度でどのように変化するのかを調べた。

1 年分の新聞コーパスを 6 カ月ごとの 2 つに分割し共起関係の解析を行う。それぞれのコーパスより得られた共起関係の情報を用いて同じ動詞間での類似度（内積値）の値を調べたとこ、図 4 に示す結果を得た。

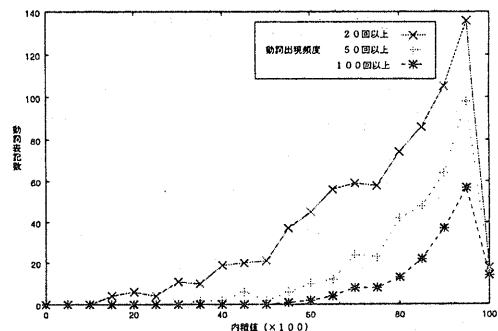


図 4: 同一動詞の内積値比較

表 3 に動詞が格助詞相当語句と共起した頻度ごとに内積値における動詞の包含率（表中の値は百分率）を表している。

内積値 (以上)	0.6	0.7	0.8	0.9
頻度 1 0 0	99.4	95.8	86.1	65.0
頻度 5 0	94.9	88.7	75.5	50.1
頻度 2 0	82.6	69.3	53.9	34.1

表 3: 同一動詞の内積比較表（包含率）

表 3 の結果によると動詞の出現頻度 1 0 0 回以上の表記では、99.4%の動詞が内積値 0.6 以上となった。また、出現頻度 5 0 回以上の

動詞についても、94.9%以上の動詞が内積値0.6以上の値をとる。

したがって、抽出元となるコーパスに偏りがなければ、表3に示した、頻度を持つ動詞を対象としてクラスタ分析を行えば、類似した格助詞相当語句のパターンを持つものをクラスタに集めることができるといえる。

3.4 分類結果

分類作業は、3.3で求めた内積値0.6をクラスタに分割する閾値として採用し、格助詞相当語句との共起頻度の合計が20以上の動詞を対象にクラスタ分析を行った。

分類結果のサンプルについては、本稿末尾に付録として表4を添付する。

3.5 分類結果に対する評価

ほとんどの動詞がクラスタに応じた概念に分類されていると思われるが、あくまで主観的なものである。

EDR概念辞書による比較を行ったが、本実験では格助詞相当語句が指し示す対象(名詞)を考慮していないため、上位から第2レベル程度の概念での共通性があるという結果に終わった。

概念の大分類として有効であるとするが、その実証については今後の課題とする。

4 考察

格助詞相当語句は、文脈上任意に用いられる格関係であるため、1つの動詞が複数の格助詞相当語句をとる。これらをまとめると次の図5に示すような関係を得ることができる。

実線で結ばれている格助詞相当語句は、クラスタを代表するものとして得られたもののうち代表的なものの関係を表現している。点線で囲んでいるものは、格助詞相当語句が持つ機能のまとまりを示したものである。

図5に示すように、格助詞相当語句の関係を整理することで格と格の間に存在している関係

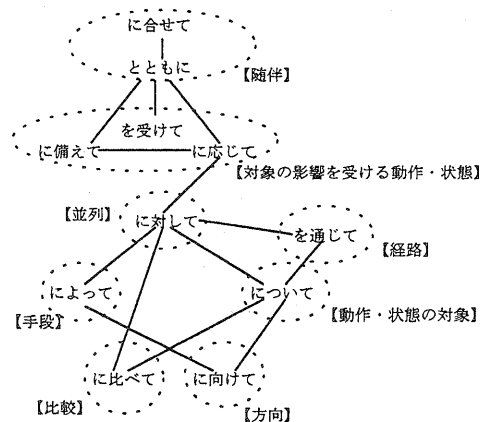


図5: 格助詞相当語句の関係

を得ることができると思われる。

5 まとめ

格助詞相当語句は文脈上任意に用いられ、どの動詞とも共起する可能性を持つため、あまり分類の重要な要素として扱われていない。一方、動詞と格助詞相当語句との関係について統計的に処理すると、動詞の持つ素性に応じて取りうる格助詞相当語句に偏りがあり、その関係は比較的安定して存在している。

格助詞相当語句は、格助詞に比べると出現する頻度が少ないため、大規模なコーパスを用いなければ十分な情報を得ることが難しい。また、文脈の影響を受けやすく、書き方・言い回しの特徴が分類結果に反映されてしまう可能性がある。

しかし、格助詞相当語句は動詞に対して限定した意味・概念関係を備えているため、得られた関係は共起している頻度が少なくても、利用価値のある情報といえる。また、格助詞相当語句は多様な表層表現を持っているため、分類の過程において人間の介入を必要としない、そのため人間の主観による影響を受けることがない。

格助詞相当語句を用いることで、動詞の持つ意味・概念関係を表層から得られた情報のみで定義することが可能であるのではないかと推察する。

本実験により得られなかった格助詞相当語句と関係を持たないから関係がないとはいえないが、動詞を代表する概念の1つのまとめ方として有効な手法であると考える。

謝辞

本研究を進めるにあたって、有意義なコメントを戴いた富士通研究所の西野文人、落谷亮の両氏に感謝するとともに、技術者派遣について、多大なるご理解とご協力戴いている、富士通株式会社の佐藤章、遠藤英明、沢辺武彦の三氏に深く感謝いたします。

また、新聞コーパスの利用を許可していただいた日本経済新聞社に感謝いたします。

参考文献

- [1] 日本電子化辞書研究所(1996). EDR電子化辞書1. 5版仕様説明書
- [2] 長尾, 佐藤, 黒橋, 角田(1996). 自然言語処理. 岩波講座, ソフトウェア科学 15
- [3] 益岡, 田窪(1997). 基礎日本語文法一改訂版一. くろしお出版
- [4] 大石, 松本(1995). 格パターン分析に基づく動詞の語彙知識獲得. 情報処理学会誌, Vol.36, No.11
- [5] 大石, 松本(1994). 格パターン分析を利用した深層格獲得手法について. 情報処理学会自然言語研究会, NL-104-10
- [6] 松本, 北内, 山下, 平野, 松田, 浅原(1999). 日本語形態素解析システム『茶筌』 version 2.0 使用説明書. NAIST Technical Report, NAIST-IS-TR99012
- [7] 張, 尾関(1997). 文節間係り受け距離の統計的性質を用いた日本語文の係り受け解析. 自然言語処理学会誌, Vol.4, No.2
- [8] 向仲(1997). 動詞と主体の属性を用いた複文の連接関係の解析. 自然言語処理学会誌, Vol.4, No.4

格助詞相当語句	動詞例	特徴
について	アドバイス, 解説, 触れる, 言及, 講演, 協議, 討議, 語り合う, 審議, 尋ねる, 議論, 検討, 質問, 話, 交換, 講義, 調査, 提言, 説明, 討論, 助言, 審査, 交渉, 聞く, 開示, 相談, 聴く, 書く, 聴取, 答申, 聞き取る, 引き続く, 詰める, 勉強, 誤る, 発言, 要望, 演説, 公開, 収集, 集計, 調整, 表示, 作成, 対処, 思い切る, 覚える, 緩和, 整理, 請求, 争う	動作・状態の対象に対するもの
について, に対し, に対して, を通じて	一貫, 注入, 割り当てる, 抛出, 供与, 融資, 補助, 支払う, 助成, 保証, 交付, 返還, 行使, 申し込む, 貸し出す, 貸し付ける, 購入, 渡す, 発注	動作・状態の対象に対し, 経路(媒体)があるもの
に応じて, に対し	支給, 選べる, 徴収, 配分, 受け取る	対象の影響を受けるもの
とともに, に応じて, に備えて	積み立てる, 預ける	対象の影響を受け, 次の事態に備えるもの
を通じて	育成, 楽しめる, 取引, 知り合う, 配布, 通報, 流入, 発行, 輸出, 蓄積, 販売, 売り出す, 流れる, 回収, 入手, 出荷, 稼働	経路(媒体)があるもの
とともに, と共に	育つ, 生きる, 成長, 搜索, 歩む, 去る, 戦う, 把握	動作の随伴があるもの
にかけて	下げ洩る, 降る, 貸す, 飛ばす	時間, 空間的な範囲に対するもの
をかけて, を投じて	改装, 建設, 増強, 整備	対象を用いるもの
とともに, に合わせて, を受けて	運行, 歌う, 企画	動作の随伴があるもの
を受けて	会見, 寄り付く, 買う, 売る	対象の影響を受けて行うもの
を経て	決定, 当選	過程をたどる時間経過があるもの
に向けて	活動, 頑張る, 協調, 出発, 前進, 動き出す, 努力, 踏み出す, 題す, 発信, 動く, 尽くす, 協力, 働きかける	方向があるもの
に比べて	下がる, 減少, 半減, 低下, 倍増, 値下がり, 向上, 悪化, 減る, 上昇, 鈍化, 好転, 値上がり, 縮小, 上積み, 増える, 増加, 劣る, 改善, 拡大, 後退, 早まる, 早める, 遅らせる, 前倒し, 値下げ, 優れる, 繰り上げる, 短縮, 落ち着く, 調達, 設定, 減額, 延長, 踏み込む, 飛ぶ, 引き上げる, 働く, 割り引く, 優遇, 出席, 選択, 増やす, 積む, 打ち上げる	比較する対象があるもの
によって, に応じて	異なる, 分かれる, 生じる, 変化, 使い分ける, 分類, 変動, 計算, 使える, 表現	対象の影響を受け, 手段があるもの

表 4: 分類結果サンプル