高頻度に共起する二用言の主格同一性の自動認識

榊 剛史　　河原 大輔　　黒橋 禎夫
東京大学工学部　東京大学大学院情報理工学系研究科
{sakaki, kawahara, kuro}@kc.t.u-tokyo.ac.jp

1 はじめに

言語処理を行うためには様々な知識が必要であるが、特に、推論や意味理解のような高度言語処理を実現するためには、因果関係のような事象と事象の関係の知識が必要となる。このような知識は、人手で書き尽くすのは非常に困難なので、大規模コーパスから自動的に抽出する研究がなされている [1, 3, 4, 5]。

本研究では、事象の単位を用言とし、高頻度に共起する二用言は意味的な関係をもつと考え、そのような用言のペアを大規模コーパスから自動的に収集する。そして、その二用言の主格が通常一致しているかどうかを自動認識し知識として蓄える。このような主格同一性の知識は、省略解析などの精度向上に役立つ重要なものである。本研究における主格同一性の認識は、高頻度に共起する二用言は同じ主格をとりやすいと考え、二用言のどちらか一方の態を変化させたときとの頻度の比較によって行う。例えば、「あげる 喜ばれる」は「あげる 喜ぶ」よりも頻度がかなり大きく、「あげる 喜ばれる」の主格は同一であると認識する。

2 用言共起ペアの収集

本研究では、まず、コーパス中で共起する二用言 (**用言共起ペア**と呼ぶ) を収集する。用言共起ペア (V_1, V_2) は、同一文内で生起し、連用修飾または連体修飾どちらかの関係をもっている二用言と定義する。ただし、用言とは、動詞、形容詞および名詞+判定詞とし、用言共起ペアは能動・受動の態の区別を含むとする。態の区別は、動詞に後続する「れる」「られる」の有無によって行う。例えば、「聞く 答える」という用言ペア*は (聞く, 答える) (聞く, 答えられる) (聞かれる, 答えられる) (聞かれる, 答える) という 4 つの用言共起ペアとしてコーパスに生起する。

用言共起ペアは以下のようにして収集する。

1. 大規模コーパスを構文・格解析する [2]。被連体修飾詞や「〜は」「〜も」などの係助詞句は、格解析により格関係を明らかにする。
2. 連用・連体修飾関係をもつ二用言を用言共起ペアとして収集する。主格同一性を認識するために、用言とともにガ格の格要素も抽出しておく。

以下では、連用・連体修飾関係それぞれの用言共起ペアについて説明する。

連用修飾関係

V_1 が V_2 を連用修飾している場合に、用言共起ペア (V_1, V_2) を収集する。次に例を挙げる。

(1) 列車に乗ってロンドンへ向かう。
→ (乗る, 向かう)

(2) 知事が辞表を提出、受理されていた。
→ (知事が 提出する, 受理される)

ただし、「燃え移る」のように 2 つの用言が隣接する場合は、複合的な 1 つの用言であることが多いので、収集する対象から除外する。

連体修飾関係

「〜V_1 した X が〜V_2 した」のように、V_1 によって修飾された体言が V_2 の格要素になっている場合に、用言共起ペア (V_1, V_2) を収集する。次に例を挙げる。

(3) 仕掛けられた爆弾が爆発する。
→ (爆弾が 仕掛けられる, 爆弾が 爆発する)

(4) 探していた楽譜を見つけた。
→ (探す, 見つける)

*コーパスにおける生起ではなく辞書的な意味において、関係する二用言を**用言ペア**と呼ぶ。

表 1: 用言共起ペアの分類

	ガ一致	出現パターン	頻度 (割合)	ガ一致評価
P1	○	Xが V_1 → Xが V_2	11,940 (0.3%)	-
P2	×	Xが V_1 → Yが V_2	662,391 (18.5%)	-
P3	○?	Xが V_1 → V_2	668,353 (18.7%)	24/30
P4	×?	V_1 → Yが V_2	855,135 (23.9%)	6/30
P4'	○?	Yが V_1 → V_2	317,516 (8.9%)	27/30
P5	○?	V_1 → V_2	1,066,132 (29.8%)	26/30

※ ただし、V_1, V_2 は 2 つの用言、X, Y は V_1 または V_2 がとるガ格である。ガ格としては、格助詞「が」で表されるガ格だけではなく、格解析の結果から分かるガ格も含む。

ただし、V_1 の連体修飾先が V_1 に対して外の関係と解析された場合は収集から除外する。これは、次の例のように、V_1 を含む節を V_2 が格要素として受けている場合であり、本研究で収集している関係とは異なるからである。

(5) 彼があいさつしたのを初めて聞いた

このようにして、新聞記事 10 年分の格解析結果から 3,581,467 個の用言共起ペアを収集した。

3 用言ペアの主格同一性の認識

用言ペアの主格一致性の認識は、態の異なる用言共起ペアを比較し、ガ格が一致する頻度が有意に高い用言共起ペアを選択することにより行う。例えば、「聞く 答える」という用言ペアは、(聞く, 答える) (聞く, 答えられる) (聞かれる, 答えられる) (聞かれる, 答える) という 4 つの用言共起ペアでコーパスに出現する。その中で (聞かれる, 答える) という用言共起ペアが、ガ格が一致して出現する頻度がもっとも高いために選択される。つまり、「聞く 答える」は (聞かれる, 答える) という形で表現されたときに主格が一致すると認識する。

以下では、まず、ガ格の一致・不一致ごとに用言共起ペアの頻度を計数することについて説明する。次に、その頻度に基づいて用言ペアの主格同一性を認識する方法について述べる。

3.1 用言共起ペアのガ格の一致・不一致

収集した用言共起ペアそれぞれにおいて、二用言のガ格の一致・不一致を調べ、その一致・不一致ごとに頻度を計数する。ガ格の一致・不一致は、二用言双方のガ格が明示されていれば明らかであるが、ガ格は省略されていることが多いので、ガ格の一致・不一致が明らかな用言共起ペアは少ない。そこで、用言共起ペアの出現パターンを分類し、表 1 のようにガ格一致の仮定をおいた。それぞれのパターンについて以下で詳しく述べる。

P1 どちらの用言のガ格も明示されていて、一致している場合。(ガ格一致=○)

(6) 彼は窃盗容疑で逮捕されたが、彼は起訴されなかった。
→ (彼が 逮捕される, 彼が 起訴される)

P2 どちらの用言のガ格も明示されているが、一致していない場合。(ガ格一致=×)

(7) ハイテク関連株が買われ、日経平均株価は二万円台を回復した。
→ (株が 買われる, 株価が 回復する)

P3 V_1 ではガ格を明示しているにもかかわらず、V_2 では省略しているので、ガ格は一致すると仮定する。(ガ格一致=○?)

(8) 寝室で夫がぐったりとしており、病院に運ばれた。
→ (夫が する, 運ばれる)

P4 P3 とは逆のパターンで、V_1 では省略しているにもかかわらず、V_2 で明示しているので、ガ格不一致と仮定する。(ガ格一致=×?)

(9) その影響を受けて、彼の才能が開花した。
→ (受ける, 才能が 開花する)

P4' P4 における Y が V_1 よりも前にあるケースで、この場合は Y が主題であることが多く、ガ格が一致すると思われる。(ガ格一致=○?)

(10) 彼は病院に収容され、すぐに死亡した。
→ (収容される, 彼が 死亡する)

P5 どちらのガ格も省略されており、二用言間でガ格は変わらないと考えられる。(ガ格一致=○?)

(11) 今日までの宿題を忘れたら、廊下に立たされてしまった。
→ (忘れる, 立たされる)

各パターンにあてはまる用言共起ペアの頻度を表 1 の「頻度 (割合)」の列に示す。

また、上記の仮定を検証するために、P3〜P5 のパターンそれぞれについて 30 個ずつ用言共起ペアを人手でチェックした。その結果を表 1 の右端に示す。それによると仮定はおおむね正しく、誤りの原因の多くは自動解析の誤りであった。

3.2 主格同一性の認識

用言ペアの主格同一性の認識は、態の異なる用言共起ペアのガ格一致・不一致の頻度を比較して、ガ格が一致する頻度が有意に高いものを選択することによって行う。用言ペアに対して、二用言の態の能動・受動の組合せ、すなわち (能動, 能動) (能動, 受動) (受動, 受動) (受動, 能動) により、4 種類の用言共起ペアが考えられる。例えば、「聞く 答える」という用言ペアの場合は、次表に示すように 4 つの用言共起ペアについてのガ格一致・不一致の頻度が得られる。

	○ or ○?	× or ×?
(聞く, 答える)	21	13
(聞かれる, 答える)	34	13
(聞かれる, 答えられる)	4	3
(聞く, 答えられる)	13	2

ただし、前節で述べたガ格一致・不一致は○, ○?, ×, ×?の 4 つに分類していたが、ここでは「○ or ○?」と「× or ×?」の 2 つに分類した頻度を用いる。

以下では、「聞く 答える」を例にとって、その主格同一性認識手法を述べる。

1. ガ格が一致する用言共起ペアの選択

   (a) 4 つの用言共起ペアから、ガ格が一致することが多いものを選ぶ。具体的には、“ガ格一致頻度/ ガ格不一致頻度” が閾値よりも高いものを選択する。この閾値は 1.6 とした。「聞く 答える」の場合、(聞かれる, 答えられる) のみ閾値を下回るので、これ以外の 3 つの用言共起ペアが選択される。

   (b) 1a で選択された用言共起ペアそれぞれの出現頻度を調べ、頻度が閾値よりも高いものを選ぶ。これは、ガ格が一致していても頻度が低く、あまり出現しない表現を選択することを防ぐためである。この閾値は、4 つの用言共起ペアの出現頻度合計の 15%とした。「聞く 答える」の場合、(聞く, 答えられる) の頻度 15 は頻度合計 103 の 14.6%であり、閾値を下回るので、これ以外の 2 つの用言共起ペアが選択される。

   ここまでの処理により、用言共起ペアが 1 つしか残っていなければ、それを選択する。1 つもなければ、その用言ペアは判別不能として処理を終える。

2. 残った用言共起ペアからガ格一致頻度が突出して高いものを選択する。残った用言共起ペアをガ格一致頻度でソートし、1 位のものと 2 位のものの頻度の差が 1.45 倍以上あるならば、1 位の用言共起ペアに決定する。「開く 答える」の場合は、(聞かれる, 答える) が 1 位 (頻度:34)、(聞く, 答える) が 2 位 (頻度:21) であり、$34/21 > 1.45$ より (聞かれる, 答える) に決定する。

各閾値は、100 個の用言ペアについてそれぞれの正解用言共起ペアを用意し、もっとも正解が多くなるような値に決定した。

上記の手順によって、132,327 個の用言ペアの主格同一性を認識した。

4 自動認識した主格同一性の評価

用言ペアの主格同一性認識結果を人手で評価した。用言ペアごとに決定した用言共起ペアの態の組合せ (能動, 能動) (能動, 受動) (受動, 受動) (受動, 能動) ごとに、ガ格一致頻度上位 50 個ずつ、合計 200 個を選び、それぞれの主格が通常一致しているかどうか人手で評価した。その評価結果を表 2 に示す。表において、○は正解、×は誤り、△は正しいが他の態の組でも主格が一致する場合があることを示す。

表 2: 主格一致性の評価結果

	○	△	×	精度 1	精度 2
(能動, 能動)	43	7	0	86.0%	100.0%
(能動, 受動)	35	7	8	70.0%	86.0%
(受動, 受動)	41	5	4	82.0%	92.0%
(受動, 能動)	40	4	6	80.0%	88.0%
計	159	23	18	79.5%	91.0%

※ 精度 1 は○のみを正解、精度 2 は○と△を正解としている。

表 2 によると、○のみを正解とした場合 (精度 1) は 79.5%、○と△を正解とした場合 (精度 2) は 91.0%であり、高い精度が得られている。表 3 に認識結果の例を示す。△と判定されたのは、例えば「わたる 行う」という用言ペアであり、(わたる, 行う) と (わたる, 行われる) のどちらでも主格が一致するからである。

主格同一性認識誤りの主な原因を以下に示す。

態の判別誤り

態の判別誤りに起因する主格同一性認識誤りがある。

(12) 9 日は、衆議院議員選挙で同日投票、即日開票される。

この例の「投票」は受動的に使われているが、「される」が後続しないため能動と誤って判別されている。このような場合、係り先の用言と態が一致することが多いので、そのような情報を利用すれば解決できると思われる。

用言共起ペアのガ格一致仮定に反する表現

用言共起ペアのガ格一致仮定が成り立たない表現に起因する主格同一性認識誤りがある。

(13) 新内閣発表を予定していたが、イスラエル軍による議長監禁策などのため延期されていた。

この文の用言共起ペアは (予定する, 延期される) なので、表 1 において P5 にあてはまり、ガ格一致と判断される。しかし、実際には「延期される」のガ格は「新内閣発表」であり、ガ格は異なっている。このような問題は、省略解析まで行うことによって解決できると思われる。

5 おわりに

本論文では、コーパスから共起する二用言を収集し、その用言のペアの主格同一性を自動認識する手法を提案した。自動認識した主格が同一となる用言ペアを人手で評価した結果、高精度に認識できていることが分かった。今後は、収集した用言ペアに関する知識を省略解析などの応用研究で活用していく予定である。

表 3: 認識結果の例

○	「運ぶ 死亡する」 → (運ばれる, 死亡する)
○	「診断する 入院する」→ (診断される, 入院する)
○	「提出する 受理する」→ (提出する, 受理される)
○	「出頭する 逮捕する」→ (出頭する, 逮捕される)
△	「リードする 迎える」→ (リードされる, 迎える)
△	「わたる 行う」 → (わたる, 行われる)
×	「投票する 開票する」→ (投票する, 開票される)
×	「予定する 延期する」→ (予定する, 延期される)

参考文献

[1] Roxana Girju and Dan Moldovan. Mining answers for causation questions. In *Proceedings of the American Association for Artificial Intelligence Spring Symposium*, pp. 15–25, 2002.

[2] Daisuke Kawahara and Sadao Kurohashi. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425–431, 2002.

[3] Christopher S.G. Khoo, Syin Chan, and Yun Niu. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 336–343, 2000.

[4] Kentaro Torisawa. An unsupervised learning method for associative relationships between verb phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 1009–1015, 2002.

[5] 乾孝司, 乾健太郎, 松本裕治. テキストから獲得可能な因果関係知識の類別およびその自動獲得の試み-接続助詞「ため」を含む文を中心に-. 言語処理学会 第 9 回年次大会発表論文集, pp. 707–710, 2003.