大規模コーパスからの名詞格フレーム構築

河野 洋志　　笹野遼平　　黒橋 禎夫
京都大学大学院情報学研究科
{kouno,ryohei,kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

計算機による文章理解に向けた重要な課題の 1 つとして、文章中の単語と単語の関係の認識がある。単語間の関係を計算機で認識するためには、人間が持っているような幅広い知識が必要であり、このような知識の代表的なものに「用言格フレーム」がある。用言格フレームとは用言の意味ごとに取り得る格の種類とその格要素の用例を記述したものであり、例えば、「積む」の場合、「荷物を積む」、「経験を積む」のように意味ごとに格フレームが存在し、格フレームごとにガ格、ヲ格などの用例が記述される。用言格フレームの例を表 1 に示す。用言格フレームをコーパスから自動構築する手法は Kawahara らにより提案されている [3]。

用言格フレームの例から、ある用言においてその意味ごとに取り得る格要素に違いがあるということがわかる。同様に名詞についても意味ごとに修飾し得る要素の違いがあると考えられる。そこで本研究では、このような名詞を修飾し得る要素のことを名詞の「格要素」、ある名詞の意味ごとにその格要素を記述したものを「名詞格フレーム」と呼ぶことにし、大規模コーパスからの自動構築を目的とする。

構築すべき名詞格フレームの例を表 2 に示す。表 2 に示した名詞格フレームから、「レバー」という名詞の場合、レバーには 2 つの意味 (『肝臓』、『操作する棒』) があり、一方の意味の場合には「牛」「生」のような名詞に修飾され、他方の意味の場合には「ブレーキ」「調整」のような名詞に修飾されるということがわかる。また逆に、ある名詞を修飾する要素を手掛りとしてその名詞の意味を解釈することもできる。例えば、「牛のレバー」という名詞句では「牛」によってレバーは『肝臓』の意味に、「ブレーキのレバー」という名詞句では「ブレーキ」によってレバーは『操作する棒』の意味であると解釈できる。

名詞格フレームは名詞句間の係り受け曖昧性解消や連想照応解析[1]などの名詞に関する様々な処理に応用することができると考えられ、文章中の単語間の関係認識を行う上で非常に重要な知識とであると言える。

表 1: 用言格フレームの例 (積む)

積む:1
ガ格　人、運転手、···
ヲ格　荷物、物、···
ニ格　車、トラック、···
···
積む:2
ガ格　人、子供、···
ヲ格　経験、体験、···
デ格　現場、会社、···
···

表 2: 名詞格フレームの例 (レバー)

レバー:1
要素 1　鶏、牛、豚、···
要素 2　生、···
···
レバー:2
要素 1　ブレーキ、アクセル、···
要素 2　手動、···
···

名詞間の関係に関する知識を集めた研究 [7][4] や既存の名詞格フレーム構築に関する Sasano らの研究 [5] は連想照応の解析を目的として名詞の必須要素に限定している。また、Sasano らの名詞格フレーム構築手法は国語辞書・シソーラスを用いた意味解析に基づいているため、未知語への対応が難しいという問題点がある。本研究ではより多様な処理に活用できるよう以下の方針に基づく名詞格フレーム構築手法を提案する。

- 修飾的な要素も含めた名詞格フレームの構築
- 辞書に依存しない手法による構築

2　提案手法

本研究では名詞 B の格フレームの構築を考えるとき、名詞句「A の B」に着目し、「A の B」とそれに共起する表現を用いて「A の B」を B の意味ごと、さらに A の役割ごとに分類することで名詞格フレームの構築を行う。提案手法の概要は以下の通りである。

[1]本研究では、先行研究 [1] にならい、「ブレーキの調子が悪いので、レバーを調整した。」という文における「ブレーキ」と「レバー」のように、照応詞と先行詞がそれぞれ別のものを指す名詞の間の照応のことを連想照応と呼ぶ。

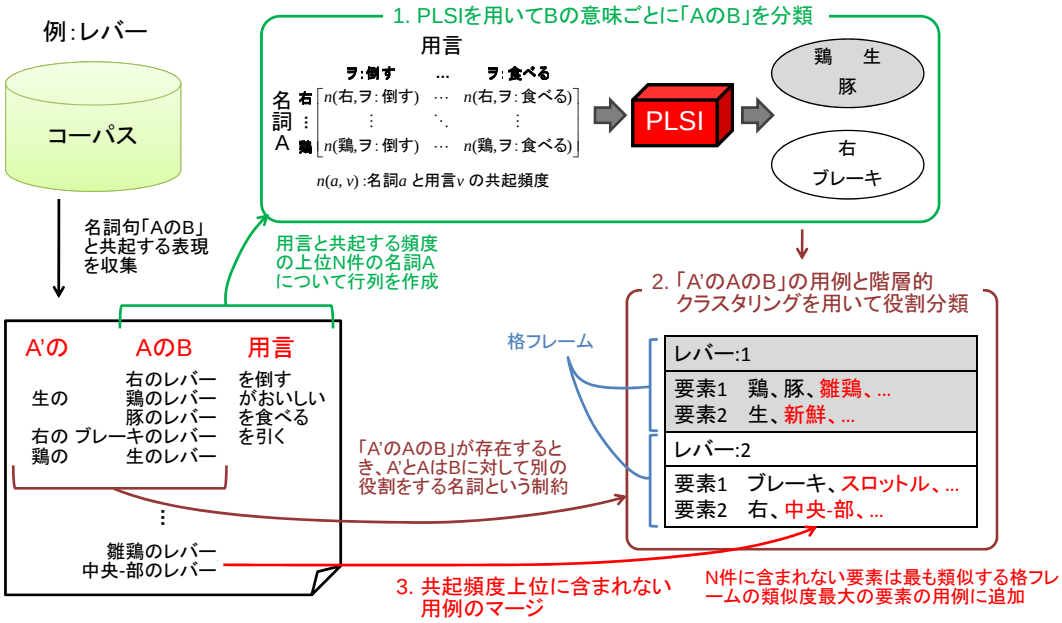

図 1: 提案手法の概要 (レバーの格フレーム)

1. PLSI を用いて B の意味ごとに「A の B」を分類
2. 「A' の A の B」の用例と階層的クラスタリングを用いて、同じ意味の B と分類された「A の B」を A の役割ごとに分類
3. 用言との共起頻度上位に含まれない用例の追加

レバーの名詞格フレームを構築する場合の概要を図 1 に示す。

2.1 格フレーム構築に用いる用例の収集

「A の B」の収集 コーパスから「A の B」の用例を収集する際は、「A の B が」や「A の B を」のように B がガ格やヲ格など A が B に係っている可能性が高い用例のみを収集する。また、B は文節内の最長の複合名詞を収集する。A は主辞を収集するが、主辞が漢字一文字の場合はその直前の名詞も併せて収集する (ex. 中央-部)。

「A の B」が係る用言の収集 「A の B」を B の意味ごとに分類するため、B が係る用言を収集する。用言の収集条件は、B がガ格、ヲ格、ニ格、デ格、ヘ格、マデ格で係る動詞、形容詞で、その格も併せて収集する。

「A' の A の B」の収集 A の役割分類に用いるため、「A の B」の直前のノ格の名詞 A' を収集する。A' には A と B の係り先曖昧性が存在するが、本研究では A' が A、B どちらに係るのかを頻度に基づいて判断し、A' が B に係っていると判断された用例のみを用いる。A' は、A と同様に主辞を収集し、主辞が漢字一文字の場合はその直前の名詞も併せて収集する。さらに、用例不足に対処するため、ある名詞句「A の B」に対して、「A' の AB」という表現が出現し、かつ A' が B に係ると解析された場合は、これらの用例を収集し「A' の A の B」の用例に追加する。

2.2 名詞格フレーム構築

前節で収集した用例を用いて名詞格フレームの構築を行う。構築は以下の 3 ステップで行う。

1.PLSI を用いた B の意味分類 PLSI[2] はソフトクラスタリング手法の一種で、確率論・情報理論に基づいて定式化されている。PLSI では潜在意味 z のもとに文書 d と単語 w が共起すると考える。このとき文書と単語の共起確率 $P(d,w)$ は

$$P(d,w) = \sum_z P(w|z)P(d|z)P(z) \qquad (1)$$

によって表される。ここで、文書 d における単語 w の出現頻度を $n(d,w)$ とすると、データの対数尤度

$$L = \sum_d \sum_w n(d,w) \log P(d,w) \qquad (2)$$

を最大にする $P(z)$、$P(d|z)$、$P(w|z)$ を EM アルゴリズムにより推定する。

一般に、PLSI は文書と単語の共起関係を扱うものであるが、本研究では名詞 B の意味によって B が係る用言の分布は異なると考え、このモデルを前節で収集した「A の B」の A と用言の共起関係について適用する。PLSI を用いて各々の A に対する B の潜在意味の帰属度を求め、帰属度が閾値 th_{PLSI} 以上の潜在意味に分類することにより、「A の B」を B の意味ごとに分類する。この際、共起する用言の種類が少ないと PLSI によるクラスタリングが上手く行えず、また

計算時間の短縮にも繋がるため、共起する用言の種類が多い上位 N_{PLSI} 件の「A の B」についてのみ PLSI の意味分類を行う。ここで、B の意味分類に PLSI に基づくソフトクラスタリングを用いるのは、B の複数の語意に対して係るような A が考えられるためである。例えば、名詞句「市販のドリル」の場合、「市販」は『工具』、『反復練習』どちらの意味の「ドリル」にも係ると考えられる。

PLSI では潜在意味の数はパラメータとして与えなければならない。本研究では潜在意味の数を 1~5 まで変化させて PLSI を実行し、モデル評価基準に赤池情報量基準 (AIC) を用いて、AIC が最小となる潜在意味数を用いる。AIC は以下の式で表される。

$$AIC = -2(L - K(N + V)) \quad (3)$$

L は対数尤度、K は潜在意味数、N は名詞 A の数、V は用言の数である。AIC はもう 1 つの代表的なモデル評価基準である BIC に比べ潜在意味数を多く見積る傾向がある。格フレーム構築の際には 1 つの語意を複数に分割しまうことよりも、別々の語意を 1 つの語意としてしまうことの方が問題となるため、AIC をモデルの評価基準に用いる。

2.「A' の A の B」の用例と階層的クラスタリングを用いた A の役割分類　「A' の A の B」という表現がコーパスに出現するとき、A' と A はそれぞれ同じ意味の B に係る要素であり、かつ B に対して別の役割を果たす要素であると考えられる。例えば、「生の牛のレバー」という用例では「生」と「牛」は共に『肝臓』の意味のレバーに係る要素であるが、それぞれ「レバー」に対して別の役割を果たしている。そこで本研究では、「A' の A の B」という表現がコーパスに出現するとき A' と A は別の格スロットである、という制約を考える。

さらに、類似した単語は同じ役割を果たすと考え、階層的クラスタリングを用いて同じ意味の B に係る A の役割分類を行う。クラスタリングに用いる類似度には、ある単語と共起する単語の分布の類似度から計算された分布類似度 [6] を用いる。階層的クラスタリング手法には群平均法 (Group Average Method) を用いる。群平均法は、任意の対象の対 x_1 と x_2 の類似度 $s(x_1, x_2)$ が与えられたデータを扱う。群平均法ではクラスタ C_1 と C_2 の類似度 $s(C_1, C_2)$ を式 (4) で計算する。

$$s(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} s(x_1, x_2) \quad (4)$$

この際、「A' の A の B」の用例が存在し、あるクラスタ (以下クラスタ 1) に名詞 A'、別のクラスタ (以下クラスタ 2) に名詞 A が含まれている場合にはクラスタ 1 とクラスタ 2 は類似度が高くてもマージしない。

階層的クラスタリングの終了条件としては、クラスタ間類似度の閾値 th_{GAM} に加え、「A' の A の B」の制約の出現割合の閾値 $\mathrm{th}_{A'AB}$ も用いる。具体的には、「A' の A の B」の制約が全体で a 個存在するとき、「A' の A の B」の制約がクラスタリングの過程で $a \times \mathrm{th}_{A'AB}$ 個出現したときクラスタリングを終了する。以下にクラスタリング終了条件をまとめる。

- 「A' の A の B」の制約の出現割合が $\mathrm{th}_{A'AB}$ 以上
- クラスタ間の類似度の最大値が th_{GAM} 以下

3. 用言との共起頻度上位に含まれない用例の追加　共起頻度上位 N_{PLSI} 件に含まれない「A の B」用例を最も類似度の高い要素の用例に追加する。

表 3: 構築された格フレーム (ドリル)

ドリル:1	
要素 1	電動:120, スレード:3, 鉄鋼用:5, ···
要素 2	市販:54, 専用:61, タイプ:21, 従来:13, ···
要素 3	径:31, サイズ:24, 太さ:14, 同径:5, ···
···	
ドリル:2	
要素 1	計算:159, 算数:167, 漢字:127, 国語:58, ···
要素 2	市販:54, 為:35, むけ:11
要素 3	大人:61, 自分:25, KUMON:3
···	

表 4: 構築された格フレーム (レバー)

レバー:1	
要素 1	鶏:189, 豚:156, 牛:78, 鳥:70, ···
要素 2	生:107, ···
要素 3	リール:38, ワイパー:55, ···
レバー:2	
要素 1	左:259, 右:260, 手元:210, 左右:141, ···
要素 2	ロック:132, スライド:60, コントロール:58, ···
要素 3	アクセル:35, 後輪錠:10, ···
···	
レバー:3	
要素 1	ブレーキ:491, スイッチ:67, ハンドル:57, ···
要素 2	切り替え:132, 調整:68, コントロール 58, ···
要素 3	小:25, 大小:16
···	

※表中の数字はコーパス中の出現頻度

3　実験・考察

2 章で述べた手法を用いてコーパス中の頻度上位 10000 件の名詞について名詞格フレーム構築を行なった。構築された格フレームの例を表 3、表 4 に示す。

3.1　PLSI を用いた B の意味分類

PLSI を用いた意味分類は用言との共起頻度上位 100 件 (N_{PLSI}=100) の A について行なった。この際、共起する名詞 A の種類が 1 つしかない用言については、十分な情報が得られず PLSI による適切な意味分類が行えなくなるとして削除した。また、名詞 A について

表 5: AIC によって選択された潜在意味数と人手による意味数の比較

名詞 B	AIC	人手	名詞 B	AIC	人手
レバー	3	2	医療	4	1
ドリル	2	2	引き出し	4	3
町長	1	1	成果	4	1
マック	2	2	レース	5	2
ツアー	5	1	定義	4	1
表情	4	1	首相	4	1
トラック	4	3	コーチ	3	1
結果	3	1	市長	4	1
チケット	3	1	達人	5	1
孫	3	1	相談	3	1

も同様に共起する用言の種類が 1 つしかないものは削除した。これを削除されなくなるまで繰り返して PLSI の入力行列とした。PLSI の帰属度の閾値 th_{PLSI} は 0.35、PLSI の EM アルゴリズムの反復回数は 300 回とした。

表 3 の例では、『工具』と『反復練習』の意味の 2 つの格フレームが構築された。PLSI によるソフトクラスタリングの結果、「市販」のようなどちらにも属すと考えられるような名詞は両方に含まれている。また、国語辞書に依存しない手法のため、「KUMON」のような未知語も名詞格フレームに含めることができ、さらに必須格に限定していないため、「専用」のような修飾的な要素も名詞格フレームに含めることができている。

PLSI によって意味分類された用例の平均の意味数は 3.56 個であった。これは先行研究 [5] に比べて多いが、AIC によるモデル評価が潜在意味を多めに見積もることの影響であると考えられる。表 5 に事前に選んだ 20 個の名詞について、AIC によって選択された潜在意味数と人手で与えた意味数の比較を示す。表 5 から、全ての名詞で AIC による潜在意味数が人手で与えた意味数を上回っていることがわかる。提案手法では基本的に複数の意味を持つと判断されているが、ほとんどの名詞には多義性はないことから、モデル評価基準についてさらに検討を行う必要がある。表 5 に示した 20 個の名詞に関しては、異なる意味の名詞が 1 つの格フレームとして構築されてしまうものは見られなかった。

次に、PLSI を用いた意味分類における誤りについて述べる。表 4 の例では、『肝臓』の意味のレバーの格フレーム [レバー:1] において「ワイパー」、「リール」が含まれてしまっている。これは以下の「ワイパー」の例のように、共起する用言のうち頻度の高い用言の『肝臓』の意味への帰属度が大きくなってしまったためである。

「ワイパーのレバー」と共起する用言の『肝臓』への帰属度　ガ:逆だ (頻度 8、帰属度 1.0)、ヲ:引く (頻度 1、帰属度 0.0)、ヲ:押す (頻度 1、帰属度 0.0)、$\cdots$

3.2 「A' の A の B」の用例と階層的クラスタリングを用いた A の役割分類

「A' の A の B」の出現割合閾値 $\mathrm{th}_{A'AB}$ は 20%とし、群平均法の閾値 th_{GAM} は 0.12 とした。

「A' の A の B」の制約の効果を確認するために、「A' の A の B」の制約を用いずに構築した場合との比較を行った。その結果、本来別の役割を持つと考えられる名詞が 1 つの要素にまとめられてしまう傾向が確認された。例えば、「ドリル」の場合、本来別の役割を持つと考えられる表 3 の [ドリル:1] の要素 1、要素 2 が 1 つの要素にまとめられてしまった。このことからこの制約が有効に働いていると言える。

4 おわりに

本稿では大規模コーパスから名詞句「A の B」とそれに共起する要素を収集し、これらを用いて名詞格フレームを構築する手法を提案した。提案手法では、PLSI による意味分類と「A' の A の B」と階層的クラスタリングを用いた役割分類の 2 段階によって、国語辞書に依存せず修飾的要素も含めた名詞格フレームを構築することができた。

今後は、構築した名詞格フレームを係り受け解析・連想照応解析に適用し、その有効性について検討していく予定である。特に、提案手法では先行研究では対象とされなかった修飾的要素も獲得しているので、名詞間の係り受け解析の精度向上が期待される。

参考文献

[1] John A. Hawkins. Definiteness and indefiniteness: A study in reference and grammaticality prediction. *Croom Helm Ltd.*, 1978.

[2] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. of SIGIR'99*, pp. 50–57, 1999.

[3] Daisuke Kawahara and Sadao Kurohashi. Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis. In *Proc. of COLING'02*, pp. 425–431, 2002.

[4] Massimo Poesio, Rahul Mehta, Axel Maroudas, and Janet Hitzeman. Learning to Resolve Bridging References. In *Proc. of ACL'04*, pp. 143–150, 2004.

[5] Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. Automatic Construction of Nominal Case Frames and its Application to Indirect Anaphora Resolution. In *Proc. of COLING'04*, pp. 1201–1207, 2004.

[6] 柴田知秀, 黒橋禎夫. 超大規模ウェブコーパスを用いた分布類似度計算. 言語処理学会 第 15 回年次大会, pp. 705–708, 2009.

[7] 村田直樹, 長尾眞. 意味的制約を用いた日本語名詞における間接照応解析. 自然言語処理, Vol. 4, No. 2, 1997.