格フレーム辞書を用いた日本語複合名詞の解析

野口慎一朗　　徳永健伸

東京工業大学 大学院情報理工学研究科

本稿では，格フレーム辞書を用いて日本語複合名詞を構成する名詞間の格関係を解析する手法について述べる．具体的には，複合名詞中のサ変名詞に着目し，サ変名詞へ係る名詞とその格関係を同定する．また，再現率を改善するために，2 種類のシソーラスを用いて名詞を拡張する手法も試みた．新聞記事から抽出した 5,000 語の複合名詞について，2 名のアノテータによって人手で解析した結果を正解データとし，提案手法を評価した．その結果，シソーラスによる拡張なしの場合で精度 76.7%，再現率 39.7%，シソーラスを用いた場合，F 値が最大の場合で精度 61.3%，再現率 55.1%の結果を得た．

Japanese compound noun analysis using case-frame information

Noguchi Sin'ichiro and Tokunaga Takenobu

Department of Computer Science
Tokyo Institute of Technology

This paper proposes a method of analysing Japanese compound nouns by using a case frame dictionary. We focus on *sahen* nouns in compound nouns and find dependent nouns on them together with the case roles. To improve recall, we used two types of thesauri to expand dependent nouns before consulting the case frame dictionary. 5,000 compound nouns were extracted from newspaper articles for evaluation. They were manually analysed by two annotators to create a reference set. The evaluation with this reference set showed the performance of 76.7% in precision and 39.7% in recall. Using the thesauri improved the result up to 61.3% in precision and 55.1% in recall at the best F-value.

1 はじめに

複合名詞の解析は自然言語処理において困難な課題のひとつである．文の統語構造を扱うための文法規則の中で複合名詞を記述しようとすると，任意個の名詞連続から名詞句を構成する規則を許すことになり，解析によって得られる統語構造の数が爆発してしまう [10]．逆にすべての複合名詞を辞書に登録することは，今日利用可能な記憶容量を考えると不可能ではないにしても，辞書の管理の点からは困難を伴う．現実的な解としては，文法を用いた統語解析に先立って，機械学習などを用いたより頑健な手法によって複合名詞を同定し，統語解析では，複合名詞をひとつの単位として扱うことが考えられる．この場合，複合名詞の内部構造の解析は別途おこなう必要がある．

複合名詞の構造解析については，これまで，意味に基づく手法 [3] ，語の共起情報に基づく手法 [6]，名詞の統語・意味的な詳細な素性に基づく手法 [9, 7] などが提案されている．語の意味 (素性) や統語的な素性を利用する手法は，その語に関する情報が得られれば精度よく解析できるが，従来，このような知識を網羅的に構築することは困難であるとされ，より単純な共起関係を使う方が網羅性という点では有利であった．しかしながら，グリット技術や記憶装置の大規模化などのおかげで，従来よりさらに大規模で，しかも，語のより詳細な素性に関する知識をほぼ自動的に構築できるようになってきた．本稿では，大量の Web ページから構築された動詞の格フレーム [4] を利用して，日本語の複合名詞の構造を解析することを試み，その結果について議論する．

2 対象データと問題設定

まず，毎日新聞 2003 年度 7 月から 12 月までの記事を形態素解析システム ChaSen [8] により形態素解析し，

(1) 名詞が 2 語以上連続し，

(2) その名詞連続の先頭以外にサ変名詞を含むものを複合名詞として抽出した．今回はサ変名詞とその他の名詞の間の格関係を中心に解析をおこなうので，条件 (2) を課しした．さらに，以下の処理を施した．

- 接頭辞，接尾辞は連接する語と合わせて 1 語とする
- 連続する数詞はまとめて一語とする

たとえば,「第二十三回」は「第/接頭辞，二/数詞，十/数詞，三/数詞，回/接尾辞」と分割されるが，上述の処理により,「第二十三回」は 1 語となる．以上の結果，34,077 語の複合名詞を抽出することができた．

表 1: 複合名詞の構成

	サ変名詞数				
名詞数	1	2	3	≥ 4	合計
2	17,905	-	-	-	17,905
3	7,865	1,868	-	-	9,733
4	2,656	981	151	-	3,788
5	895	510	108	8	1,521
≥ 6	648	317	144	21	1,130
合計	29,969	3,676	403	29	34,077

複合名詞を構成する名詞数，サ変名詞数で分類した結果，表 1 のような分布になった．表 1 より，名詞 2 語からなる複合名詞が 50%以上を占め，またサ変名詞を 1 つしか含まない複合名詞が 90%近くを占めていることがわかる．

このような前処理を施すと任意の複合名詞は，単語列 $n_1, n_2, \ldots n_k$ で表現できる．ただし，n_i は接頭辞，接尾辞を含む場合もある．また，この中には最低でも 1 つのサ変名詞を含む．サ変名詞を n_i^* で表現すると，本稿で扱う問題は，k 語からなる単語列 $n_1, n_2, \ldots n_k$ の各単語に格と係り先のサ変名詞の組 (c, n_i^*) を付与することとなる．

3 格フレーム辞書よる格関係の解析

以下の手順にしたがって，Web ページから構築された京都大学で構築された格フレーム辞書 [5] を用いて，複合名詞を構成する各名詞とその係り先の名詞の格関係を解析する．この格フレーム辞書では，動詞を検索キーとして，その動詞

に係る名詞と格関係を頻度と共に検索することができる．

(1) 複合名詞中からサ変名詞 (係り先) を見つける．

(2) 格フレーム辞書を用いて，(1) で見つけたサ変名詞を検索し,「係り要素」「格関係」を取得する．

(3) サ変名詞より左に出現する名詞と，(2) で取得した「係り要素」のマッチングをおこない，マッチしたものを格関係があると認定する．

今回は，可能性をすべて列挙しており，ひとつの候補に絞り込むことはしていない．また，(3) で複合名詞中の名詞をマッチングするときに，名詞が接頭辞あるいは接尾辞を含む場合は，最後部の要素を用いる．

4 正解データの作成

4.1 人手による格関係の付与

前節の手法を評価するために，正解データを作成する．まず，新聞記事から抽出した 34,077 の複合名詞から，表 1 の分布を反映するように 5,000 語をランダムに抽出した．ただし，名詞数, サ変名詞数が多いものは表 1 の分布より多めに抽出した．この 5,000 語のうちランダムに選択した 1,000 を 2 名のアノテータに独立に解析させた．また残りの 4,000 語は 2,000 語ずつを 2 名のアノテータに解析させた．格関係の認定にあたっては以下の指示を与えた．

1. 係り要素は係り先に先行する．
2. 「ノ格」は使わず，連体修飾関係には記号 “M” を記入する．
   (例) 「改訂/作業」…「改訂の作業」なので格関係は “M” とする．
3. 係り先が複数考えられる場合はすべて付与する．
   (例)「イラク/占領/統治」…「イラクを占領する」,「イラクを統治する」
4. 格関係が複数考えられる場合はすべて付与する．
   (例)「海外/出張」…「海外に出張する」,「海外へ出張する」
5. 基本的に能動態の場合を考え，受動態，使役は使わない．
   (例) 「局面/打開」…「局面を打開する」
6. 並列はそれぞれに格関係を付与するが，並列なものが同時にあるアクションを行う場合は「ト格」を付与する．
   (例) 「電気/空調/検査」…「電気を検査する」,「空調を検査する」
   「日/米/貿易」…「日と米が貿易する」(「日が貿易する」,「米が貿易する」とは分けない).

表 2: 正解データ 5,000 語の分布

	サ変名詞数				
名詞数	1	2	3	≥ 4	合計
2	2,705	-	-	-	2,705
3	1,180	282	-	-	1,462
4	404	143	26	-	573
5	104	68	9	3	184
≥ 6	42	18	12	4	76
合計	4,435	511	47	7	5,000

4.2 解析結果の一致率

アノテータ 2 名の解析結果をそれぞれ A，B とし，共通部分 1,000 語の解析結果を比較したものを表 3 に示す．

表 3: 解析結果の一致率

A が解析した名詞数	1,534
B が解析した名詞数	1,553
完全に一致した数	1,156
一部分が一致した数	58
格関係のみが異なる数	250
係り先のみが異なる数	114

表 3 より，解析結果が一部分でも一致する場合は約 80%程度であること，結果が異なる場合の約 73%は格関係が異なっていることがわかる．このデータの κ 統計量は 0.78 であった．格関係

が異なる場合を詳しく見てみると表 4 のような結果となった．付与された格の大部分は「ガ格」,「ヲ格」,「ニ格」,「デ格」，“M” であり，その他は全体の 1 割に満たなかった．基本的には連体修飾関係 “M” の部分を間違えていることが多い．

表 4: 格関係の交差行列

A ＼ B	ガ	ヲ	ニ	デ	M
ガ	174	10	5	6	9
ヲ	19	401	3	12	18
ニ	8	12	168	15	8
デ	14	10	5	132	13
M	7	24	17	20	260

4.3 正解データの作成

共通部分 1,000 語については，両者の結果の和集合を正解データとした．また，今回用いた解析手法は，連体修飾関係やサ変名詞以外に係る格関係については解析の対象としていないので，正解データの中から除外した．その結果，複合名詞 5,000 語中で，係り先のサ変名詞とその格関係の組が認定できた名詞は 4,125 であった．これらの名詞に対して付与した係り先と格関係の組が評価の対象となる．

5 解析手法の評価

正解データを得た複合名詞 5,000 語について，3 節で提案した手法で解析をおこなった．解析結果，正解データ共に複数の解がありうるので，各名詞について両者に共通な組が 1 つでもあれば正しいものとする．評価結果を表 5 に示す．表 5 より，精度に比べて再現率が低いことがわかる．これは格フレーム辞書の網羅性が不十分であることを示している．

表 5: 評価結果

解析対象の総名詞数	4,125
解析できた名詞数	2,133
正しく解析できた名詞数	1,636
再現率 [%]	39.7
精度 [%]	76.7

6 解析手法の拡張

6.1 拡張手法の提案

再現率を改善するために，複合名詞中に含まれる名詞と格パターン辞書から得られる情報をマッチングする際に，複合名詞中の名詞をシソーラスを用いて拡張することを試みた．シソーラスとして Digital Library [1] のシソーラス検索辞書と分類語彙表 [2] の 2 種類を用いた．Digital Library のシソーラスでは，同義語，狭義語，広義語の関係を用いて名詞を拡張する．分類語彙表は同じカテゴリーに属する語を用いて名詞を拡張する．

係り元，係り先 (サ変名詞) の名詞の拡張は表 6 の 5 種類の拡張パターンを考え，このパターンを組み合わせることで，複数の拡張手法を考える．ここで “DL”，“BH” はそれぞれ Digital Libarary (DL)，分類語彙表 (BH) を使って名詞を拡張することを，“–” は拡張しないことを意味する．原理的には 9 通り (3×3) の組み合わせが考えられるが，分類語彙表の同一カテゴリ内の語の関係は多様なので，係り先 (サ変名詞) の拡張には用いなかった．

表 6: 名詞の拡張パターン

拡張パターン	係り元	係り先
Ex1	DL	–
Ex2	BH	–
Ex3	–	DL
Ex4	DL	DL
Ex5	BH	DL

これらの拡張パターンを組み合わせて，表 7 の 10 種類の拡張手法について実験を行った．表中の「→」は解析結果が出なかったら先に進み，結果が出たらそこで終了することを表し,「+」は 2 つの結果を同時に使うことを表している．

係り元の名詞を先に拡張するのは，係り先の名詞を拡張すると格関係が大幅に変化する可能性があるためである．この拡張手法は 3 章の手法で解析できなかった場合についてのみ適用する．

表 7: 拡張手法

拡張手法	拡張パターンの適用順序
手法 1	Ex1
手法 2	Ex2
手法 3	Ex1 → Ex2
手法 4	Ex2 → Ex1
手法 5	Ex1+Ex2
手法 6	Ex1 → Ex3 → Ex4
手法 7	Ex2 → Ex3 → Ex5
手法 8	Ex1 → Ex2 → Ex3 → Ex4 → Ex5
手法 9	Ex2 → Ex1 → Ex3 → Ex5 → Ex4
手法 10	(Ex1+Ex2) → Ex3 → (Ex4+Ex5)

6.2 解析結果の比較

手法ごとの解析結果を比較するために再現率，精度，F 値，解析時間をそれぞれの手法ごとに計算した結果を表 8 に示す．F 値は再現率と精度から以下の式で計算する．解析時間は 1 語を解析するのにかかった平均解析時間を表している．

$$\text{F 値} = \frac{2 \times \text{再現率} \times \text{精度}}{\text{再現率} + \text{精度}}$$

表 8: 拡張手法ごとの解析結果の比較

拡張手法	再現率 [%]	精度 [%]	F 値	解析時間 [秒]
拡張なし	39.7	**76.7**	52.3	**0.005**
手法 1	45.4	68.8	54.7	0.015
手法 2	53.7	61.9	57.5	0.34
手法 3	54.6	60.7	57.5	0.32
手法 4	54.6	60.8	57.5	0.40
手法 5	55.1	61.3	**58.1**	0.41
手法 6	48.4	62.6	54.6	0.44
手法 7	57.4	56.5	56.9	1.22
手法 8	57.9	55.5	56.7	1.16
手法 9	58.0	55.6	56.7	1.26
手法 10	**58.5**	56.1	57.3	1.32

6.3 結果に関する考察

表 8 から，名詞を拡張することによって再現率が上昇していることがわかる．手法 1 と 2，および手法 6 と 7 の比較から，Digital Library と分類語彙表では，分類語彙表を用いた方が再現率が大幅に上昇していることがわかる．手法 3 と 4 と 5，8 と 9 と 10 の違いは，拡張の結果において，Digital Library を優先するか，分類語彙表を優先するか，両者とも同等と見なすかの違いである．表 8 からは Digital Library と分類語彙表の両方を使う場合はどちらを優先しても結果に違いはほとんど見られないことがわかる．

係り先の名詞の拡張を行わなかった場合と行った場合（手法 3 と 8，手法 4 と 9，手法 5 と 10）を比較すると，係り先を拡張しなかったた方が F 値の値が大きいことがわかる．これは再現率の上昇に比べ，精度の低下が大きいためである．

6.4 誤り分析

解析結果の誤りは主に次の 3 種類に分類できる．

- 存在しない格関係を認定している (ノイズ)
- 認定すべき格関係が認定できない (漏れ)
- 係り先あるいは格関係が誤っている (エラー)

ノイズは本来は連体修飾をしている名詞に格関係を認定している場合が多い．漏れはサ変名詞が格フレーム中に存在しないものと，格フレーム中に係り元の名詞が存在しないものの 2 つの場合があるが，後者の場合がほとんどで，固有名詞を含むものが多かった．エラーは名詞をシソーラスで拡張した結果生じることが多かった．他に間違った原因として見られたのは，形態素解析による区切り位置が間違っている場合で，特にカタカナ語表記のものに多く見られた．

手法ごとに間違いのタイプを調べた結果を表 9 に示す．表 9 より，拡張を行わない場合はノイズとエラーが少ないが，漏れがかなり多く，拡張を行うと漏れが減少する代わりにノイズとエラーが増えていることがわかる．

7 まとめ

本稿では，格フレーム辞書を用いた複合名詞中の名詞間の格関係の解析方法を提案した．新聞記事から抽出した複合名詞を 2 名で解析し，これを正解データとして，提案手法を評価した結果，精度 76.7%，再現率 39.7%という性能が得られた．再現率が低いのは格フレーム辞書の網羅性が十分ではないことを示唆している．この問

表 9: 拡張手法ごとの誤りタイプ

拡張手法	ノイズ	漏れ	エラー
拡張なし	375	2,362	190
手法 1	658	2,031	292
手法 2	1,101	1,571	419
手法 3	1,175	1,512	445
手法 4	1,175	1,512	443
手法 5	1,175	1,512	421
手法 6	946	1,827	374
手法 7	1,497	1,324	525
手法 8	1,574	1,278	548
手法 9	1,574	1,278	543
手法 10	1,574	1,278	523

題を解決するために，2 種類のシソーラスを用いて，いくつかの組み合わせで複合名詞中の名詞を拡張する実験をおこなった．その結果，再現率は最大で 58.5%まで改善できることがわかった．

今回は抽出されたすべての複合名詞を解析の対象としたが，固有名詞を含むものは格フレーム辞書の検索に失敗することが多く，これが全体の性能を下げる原因になっていると考えられる．現在の格フレーム辞書の網羅性を正確に調べるためには，固有名詞を含む複合名詞を除外して考える必要がある．固有名詞を含む複合名詞を解析するためには，まず，固有名抽出の技術を使い，複合名詞中の固有名詞のタイプを同定し，それを手がかかりに格フレーム辞書を検索するなどの前処理が必要である．

今回は格関係のみを解析の対象としたが，人手で解析したデータから連体修飾関係も多く含まれることがわかっている．今後は連体修飾関係の同定も考慮する必要がある．

最後に，今回は可能な格関係をすべて列挙するだけで，解析結果をひとつに絞り込むことはしなかった．同じサ変名詞に係る他の格要素も考慮することによって，あるいは格フレーム辞書の頻度情報などを利用して，解を一意に絞り込むことも必要である．

参考文献

[1] Digital library －シソーラス辞書検索サービス. http://digilib.silkroad.net/modules/.

[2] 『分類語彙表 –増補改訂版–』データベース (version 1.0).

[3] 横山晶一, 佐久間一弘. 意味素性を用いた複合名詞の生成による分析. 計量国語学, Vol. 20, No. 7, pp. 304–314, 1996.

[4] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた web からの大規模格フレーム構築. 自然言語処理研究会, Vol. 171-12, pp. 67–77, 2006.

[5] 京都大学黒橋研究室. 格フレーム検索サービス. http://reed.kuee.kyoto-u.ac.jp/cf-search/.

[6] 小林義行, 徳永健伸, 田中穂積. 名詞間の意味的共起情報を用いた複合名詞の解析. 自然言語処理, Vol. 3, No. 1, pp. 29–43, 1996.

[7] 竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫. 語彙概念構造を利用した複合名詞内の係り関係の解析. 情報処理学会論文誌, Vol. 43, No. 5, pp. 1446–1456, 2002.

[8] 奈良先端科学技術大学院大学. 日本語形態素解析システム ChaSen「茶筅」ver.2.3.3. http://chasen.naist.jp/hiki/ChaSen/.

[9] 内山清子, 竹内孔一, 吉岡真治, 影浦峡, 小山照夫. 専門分野における複合名詞解析のための名詞文法属性の分類について. 計量国語学, Vol. 23, No. 1, pp. 1–24, 2001.

[10] 野呂智哉, 橋本泰一, 徳永健伸, 田中穂積. 大規模日本語文法の開発. 自然言語処理, Vol. 12, No. 1, pp. 3–32, 2005.