全文検索エンジンを用いた単なる名詞列からの概念具体物関係の抽出

隅田 飛鳥，鳥澤 健太郎，新里 圭司
北陸先端科学技術大学院大学情報科学研究科
Email: {a-sumida，torisawa，skeiji}@jaist.ac.jp

1 はじめに

語彙統語パターン (lexico-syntactic patterns) は，ある特定の表現から上位下位関係や，概念具体物関係 (concept-instance relation, 上位下位関係の一部とみなせ，例えば概念名 “映画” に対する具体物名 “キングコング” などが考えられる) を得るために用いられてきた [1, 2, 3, 4, 5, 6, 7]．しかし語彙統語パターンは限定された表現を対象として上位下位関係や概念具体物関係を獲得するため，情報検索や情報抽出などの応用分野で必要となる広範な関係を得るのが難しい．そのため，幅広く上位下位関係や概念具体物関係を獲得するには，より多様な表現から関係を獲得する必要がある．

本稿では，語彙統語パターンを適用できない日本語の単なる名詞列から大量の概念具体物関係を獲得することを目指す．日本語の名詞列は，概念具体物関係を示す二つの名詞列の並置であることがある．概念具体物関係を示す名詞列の例を図 1 に示す．図 1 中の タイプ A は概念名と具体物名の境界が括弧によって明示された名詞列，タイプ B は概念名と具体物名の境界を示す手掛りがない名詞列を表す．以下，本稿ではタイプ A を**括弧つき名詞列**，タイプ B を**括弧なし名詞列**と呼ぶ．二つのタイプのうち，括弧つき名詞列からは括弧を用いた単純なパターンを適用することで概念具体物関係を獲得できる．しかしながら, 括弧なし名詞列については概念名と具体物名の境界を確実に識別する方法が存在しないため，単純なパターンでは概念具体物関係を獲得できない．英語の場合，固有名詞の先頭は大文字になるなどの手がかりを用いることでタイプ B の名詞列からでも比較的容易に概念名と具体物名の境界を識別できる．しかし日本語の場合は，境界を示す明確な手がかりがないため，英語の場合と比べタイプ B の名詞列から境界を識別することは困難である．

本研究では，括弧なし名詞列から概念具体物関係を獲得することを目的とする．括弧なし名詞列から概念具体物関係を比較的高い精度で獲得するために，括弧つき名詞列から獲得した関係を利用する．本手法では，誤って獲得された概念具体物関係を除くため，WWW 文書集合を対象とする全文検索エンジンを用い，フィルタリングを行う．また本手法の副産物として，括弧なし名詞列から得られた概念具体物関係以外にも，その途中段階で括弧つき名詞列から非常に多くの高精度な概念具体物関係を獲得できる．また，本稿では日本語を対象に設計したが，中国語などの他言語についても単なる名詞列に関して同様の手法が適用可能であると考えている．

タイプ A：括弧つき名詞列
映画「 比丘尼 物語 」
概念名：映画，**具体物名**：比丘尼 物語
描画 ソフト「 シルエット 」
概念名：描画 ソフト，**具体物名**：シルエット
タイプ B：括弧なし名詞列
宿 丸栄
概念名：宿，**具体物名**：丸栄
仮想 商店 街 楽天 市場
概念名：仮想 商店 街，**具体物名**：楽天 市場

図 1: 概念具体物関係を示す名詞列の例

概念具体物関係を獲得する既存手法として，英語を対象とした Fleischman らの手法 [2] があるが，本手法には以下のような違いがある．(1)Fleischman らが用いているコンマや固有名詞の先頭は大文字であるなどの明確な手がかりが日本語では存在しないため，このような手がかりに頼らない．(2)Fleischman らが特定の分野に依存した素性を用いた機械学習により関係をフィルタリングするのに対し，本手法では特定分野の知識に依存しないフィルタリングを行う．

2 提案手法

括弧なし名詞列から概念具体物関係を獲得するためには，まず括弧なし名詞列が概念具体物関係を示す名詞列かどうかを識別する必要がある．そこで，括弧あり名詞列から獲得される概念具体物関係中に頻出する概念名を選び出し，それら概念名で始まる括弧なし名詞列のみを概念具体物関係を示す名詞列であるとする．これは，**概念名を示しやすい名詞列が存在し，この名詞列から始まる括弧なし名詞列は概念具体物関係を示すことが多い**，という我々の直観に基づく．この直観に従い，本手法は以下の 2 ステップに分けて概念具体物関係を獲得する．

ステップ 1 括弧つき名詞列からの概念名の抽出

ステップ 2 全文検索エンジンを用いた，括弧なし名詞列からの概念具体物関係の抽出

まず，ステップ 1 で括弧つき名詞列より獲得された概念具体物関係中に頻出する概念名を抽出する．次にステップ 2 では，ステップ 1 で抽出された概念名から始まる括弧なし名詞列から概念具体物関係候補を抽出し，さらに全文検索エンジンより得られる統計値を利用してフィルタリングし，最終的な概念具体物関係を得る．

以下，各ステップについて詳しく述べる．

2.1 ステップ 1 : 括弧つき名詞列からの概念名の抽出

ステップ 1 の目的は，括弧なし名詞列から概念具体物関係を獲得するための準備として，括弧つき名詞列から概念名候補集合を抽出することである．

括弧つき名詞列 (図 1 のタイプ A) は，以下のように表現される．

$$\underbrace{N_1 N_2 \cdots N_{i-1}}_{\text{概念名候補}} \underbrace{「 N_i N_{i+1} \cdots N_m 」}_{\text{具体物名候補}}$$

ここで，単語 N_i は形態素解析器が名詞または未知語と判断した単語である．手続きとしては， まず， コーパスを形態素解析 (MeCab[1]を利用) して， 以上のパターンを持つ括弧つき名詞列を抽出し， 概念具体物関係候補集合とする．ここで，予備実験で得られた知見から，概念名候補が接尾詞, 数, 固有名詞, 機能語で構成されている場合と, 具体物名候補が接尾詞, 機能語のー単語のみで構成されている場合について, 概念具体物関係候補から取り除いた．ここまでの手順を HTML 文書集合 13.1GB に適用した結果，69,665 個の概念具体物関係候補が得られた．獲得した概念具体物関係候補からランダムに 100 個選び出し評価した結果，その適合率は 58%であった．

ついで，得られた概念具体物関係候補から概念名候補のみを抽出し，概念名候補集合とするが，このままではもととなる概念具体物関係候補の適合率が 50% 台と低いため，以下のような作業を行う．まず，予備実験では，概念具体物関係候補が正しく関係を示しているかどうかは，概念名候補に大きく依存しており，特に概念名候補中の主辞となる名詞 (上述のパターンでは，概念名候補の末尾の名詞 N_{i-1} に対応) への依存度が高いことが分かった．以下では概念名候補中の主辞を**概念名候補主辞名詞** と呼ぶ．そこで，適切な概念具体物関係を示しやすい概念名候補を以下の手順によってリストアップした．まず，概念名候補主辞名詞を，同一の概念名候補主辞名詞をもつ概念具体物関係候補の異なり数に従い，ソートする．ここで，ソートした概念名候補主辞名詞で最も多くの概念具体物関係候補に現れている 409 語は，概念具体物関係候補全体の約 60%に現れていることが分かった．この 409 語の概念名候補主辞名詞は 13,381 個の概念名候補の主辞となっていた．

ここで，以下のような手順によって，適切な概念具体物関係を示しやすい概念名候補主辞名詞を先の 409 個の概念名候補主辞名詞から選び出した．まず，先の 409 個の概念名候補主辞名詞各々に対して，それを含む概念具体物関係候補をランダムに 5 つ抽出した．ついで，概念具体物関係候補が適切であるかどうかを人手でチェックした． もしチェックした概念具体物関係候補中に 3 つ以上誤りがあれば，概念名候補から，概念名候補主辞名詞を主辞名詞としてもつものすべてを取り除いた．この手作業によるフィルタリングにより，301 個の概念名候補主辞名詞が得られた．さらに，得られた概念名候補主辞名詞をもつ概念名候補の集合から，あまり意味のない概念名 “名”, “物”, “名前” を取り除き，最終的に概念名候補 11,154 個が得られた．この概念名候補集合はステップ 2 で利用する．

手動で概念名のフィルタリングを行った後，得られた概念名候補集合を用いて，括弧つき名詞列から獲得された 69,665 個の概念具体物関係候補を制限した．これにより，28,135 個の関係候補に絞りこむことができた．この絞り込まれた概念具体物関係候補からランダムに 100 個の関係を選び出し評価した結果，概念名を制限する前と比べ，適合率は 89%に向上した．この結果から，概念名候補主辞名詞を対象とした手動フィルタリングをすることで適合率を大幅に改善でき，括弧つき名詞列から非常に正確な概念具体物関係を獲得できることを確認できた．

もちろん，概念名候補集合の要素数は多いほうが有益だが，概念名候補のチェックを人手で行うコストとデータスパースネスにより，概念名候補の数は限られる．仮りに，より巨大なウェブ文書集合を対象にチェックを行い，多くの時間をかけることができれば，概念名候補集合の要素数の制限は大幅に改善するだろうと考えられる．

括弧つき名詞列から上位下位関係を獲得する手法は今角により提案されている [7]．今角によれば，本手法で行っている上位語の手動チェックをすることなしに 90%以上の適合率で上位下位関係が獲得できたと報告されている．本手法でも，今角と同様に括弧つき名詞列から概念具体物関係の獲得を行ったが，その適合率は 89%であった．この適合率が下がった原因は，本手法では HTML 文書をコーパスとして用いているのに対し，今角は新聞記事をコーパスとして利用しているためだと考えられる．新聞記事は訓練された書き手により書かれるため，本研究で対象としている HTML 文書と比べ全体的に多様性が小さい．

2.2 ステップ 2 : 全文検索エンジンを用いた，括弧なし名詞列からの概念具体物関係の抽出

ステップ 2 は以下に示す 4 つの手順からなる．

ステップ 2.A WWW 文書集合からの名詞列の抽出

ステップ 2.B ステップ 1 で獲得した概念名候補集合を用いた名詞列のフィルタリング, 及び概念名・具体物名間の境界決定

ステップ 2.C 品詞による名詞列のフィルタリング

ステップ 2.D 全文検索エンジンと概念名の長さによるフィルタリング

ステップ 2.A では，単純に WWW 文書集合から名詞と未知語の連続を名詞列として抽出する．次にステップ 2.B で，ステップ 1 で獲得した概念名候補集合の各要素が先頭の部分名詞列と一致する名詞列を抽出し，抽出された名詞列に含まれる概念名候補の終端を概念名候補と具体物名候補との境界と仮定する．本研究では，問題を単純化するために，具体物名候補を 1〜2 単語で構成される単語に限定する．続いてステップ 2.C では，予備実験より，具体物名候補中に接尾詞, サ変名詞を含む場合は, 適切な概念具体物関係を示しにく

[1] http://chasen.org/~ taku/software/mecab/

いことが分かったため，このような概念具体物関係候補を取り除く．

ステップ 2.C の出力結果は以下のパターンを持つ．

$$\underbrace{C_1C_2\cdots C_i}_{\text{概念名候補}}\ \underbrace{I_1\{I_2\}}_{\text{具体物名候補}}$$

但し，C_i，I_j はそれぞれ名詞もしくは未知語を示す．

ステップ 2.D では，以下の 4 つのヒューリスティックルールに従い，WWW 文書集合を対象とした全文検索エンジンから得たヒットカウントと概念名候補の長さを利用したフィルタリングを行う．以下，$hit(query)$ は文字列 $query$ を検索エンジンに与えたときのヒットカウントを示す．

ルール 1 単語 “どの” を具体物名候補に付け加え，検索エンジンに検索語として与える．hit(“どの $<$ 具体物名候補 $>$ ”) が閾値 θ より大きければ，この具体物名候補をもつ名詞列を概念具体物関係候補集合から取り除く．このルールは，**具体物名は固有名詞として機能するため，検索語 “どの ＜具体物名候補＞ ” は WWW 文書集合に現れにくいだろう**，という直観に基づく．ヒットカウントが多ければ，具体物名候補は一般的な概念を表す語とみなせるため具体物名ではないと考えられる．

ルール 2 名詞列が概念名候補 $C_1C_2\cdots C_i$，具体物名候補が 2 単語 I_1I_2 で構成されていると仮定する．$hit(C_1\cdots C_iI_1)/hit(C_1\cdots C_iI_1I_2)$ が閾値 σ より大きければ，概念具体物関係候補集合からこの名詞列を取り除く．このルールは，$C_1\cdots C_iI_1$ **のヒットカウントが** $C_1\cdots C_iI_1I_2$ **と比較して大きければ，名詞列の主要な意味的分割が，概念名候補** $C_1\cdots C_i$ **と具体物名候補** I_1I_2 **の間ではなく，** $C_1\cdots C_iI_1$ **と** I_2 **の間に存在し，** $C_1\cdots C_i$ **と** I_1I_2 **とが適切な概念具体物関係を構成しないであろう**という直感に基いている．より詳細な説明は後述する．なお，具体物名候補が 1 単語で構成されている場合は，このルールを適用しない．

ルール 3 hit(具体物名候補) が閾値 ξ より大きければ，概念具体物関係候補集合からこの具体物名候補をもつ概念具体物関係候補を取り除く．このルールは，予備実験により得られた**具体物名候補が WWW 文書集合中に頻出するならば，一般的な概念名としてみなせる**という観察結果に基づく．

ルール 4 もし，概念名候補が 1 文字ならば，概念具体物関係候補集合からこの概念名候補をもつ名詞列を取り除く．日本語を対象とした形態素解析器は，長い固有名詞の認識を誤り，1 文字の単語列として分解することがある．このような単語列の先頭に位置する単語を概念名候補として仮定した場合，この単語列を概念具体物関係候補集合から取り除く．

ステップ 2.D では，上記のヒューリスティックルールをルール 1 から 4 の順に適用する．本研究では，各パラメータの値として予備実験の結果より決定した $\theta = 0$，$\sigma = 1.5$，$\xi = 2\times10^4$ を用いている．また，ヒットカウントを得るための検索エンジンとしては，あらかじめ WWW より収集した 0.7TB の HTML 文書を検索対象とする独自のものを用いている．

ルール 2 について補足する．例として 名詞列 “歌劇 場 管弦楽団” を考える．ここでは仮に，ステップ 1 で “歌劇” が概念名候補として獲得されているとする．この場合，“場 管弦楽団” は概念名候補 “歌劇” の具体物名候補だと考えられるが，この概念具体物関係候補は誤りである．しかし，この概念具体物関係候補はルール 1，3，4 を適用しても関係候補集合から取り除かれない．例えば，ルール 1 で “どの 場 管弦楽団” と検索しても，そもそもこのような複合名詞がまれにしか出現しないため，具体物名候補を一般的な概念として認識できないためである．

ルール 2 は，このような境界検出誤りを検出し，取り除くために利用する．hit(歌劇　場) が，hit(歌劇　場　管弦楽団) と比較し十分大きいならば，“歌劇 場” は独立して出現する複合名詞とみなせる．つまり，“場” と “管弦楽団” 間の境界であり，“歌劇” と “場” 間の境界 でないことを示す．

ステップ 2 の説明は以上である．本手法の最終的な出力結果は，概念具体物関係候補の集合に上記のヒューリスティックルールを適用した後の概念具体物関係集合である．

3　実験

まず，WWW からダウンロードした文書集合 0.7TB を対象に全文検索エンジンを構築した．WWW 文書集合からランダムに選んだ未知の HTML 文書群 6.5GB に対し，本手法を適用した．その結果を表 1 に示す．表 1 の各列は左から順に，**A)** 各ステップにおける概念具体物関係の数，**B)** 各ステップにおける概念具体物関係の適合率 (獲得された関係からランダムに 100 個抽出し評価)，**C)** B) の適合率より求められる正しい概念具体物関係と期待できる関係の数を示す．但し，表中の各値は概念具体物関係候補から重複を除いた値である．本手法で得られた概念具体物関係候補は数が多く，その再現率を求めることは困難である．そのため本稿では，再現率の代替尺度として，正しい概念具体物関係を得ることが期待できる概念具体物関係の数を用いた．表より，ステップを経ることで適合率が向上していることが分かる．また表には，Step 2.D の各ヒューリスティックルールを適用した結果と各ルールがどの程度適合率の向上に貢献しているかも示してある．各ルールの適合率は，その都度ランダムに抽出した 200 個の概念具体物関係から求めた．各ルールを適用していくことで適合率が向上することが表より分かる．

本手法のステップ 1，2 を適用することで，最終的に得られた概念具体物関係の数は 4,276 個であり，その適合率は 83.5%であった．獲得された 4,276 個の概念具体物関係の中には 685 個の概念名候補が含まれていた．実際に獲得できた概念具体物関係の例を図 2 に示す．

次に，同じ文書集合に含まれる括弧つき名詞列に対し，ステップ 1 で得られた概念名候補集合を用いて概念名に制限を加えた場合 (ステップ 1 に相当) とそうでない場合の両方について，獲得された概念具体物関係

表 1: 各ステップ・ヒューリスティックルールによって得られた結果

ステップ	概念具体物関係の数	適合率 [%]	予測される正しい関係の数
ステップ 2.A	4,107,460	-	-
ステップ 2.B	117,443	10	11,744
ステップ 2.C	58,322	19	11,081
ステップ 2.D	4,276	83.5	3,570
ルール 1	14,418	54	7,785
ルール 2	6,140	66.0	4,052
ルール 3	5,337	75.5	4,029
ルール 4	4,276	83.5	3,570

株式会社 / 千馬	哲学 者 / セネカ
医療 法人 / 慈誠会	クラブ / 櫻 セレブリティ
景勝 地 / 蘇 洞門	小説 / 吉田 学校
絵本 / 即興 詩人	
前奏曲 / ト長調*	ＣＤ / ロムロム*

“/” は本手法により検出された概念名候補と具体物名候補との境界
“*” が末尾についている概念具体物関係は正しくない関係を表す.

図 2: 獲得された概念具体物関係の例

を評価した．その結果を表 2 に示す．但し，表中の各値は概念具体物関係候補から重複を削った後の値である．括弧つき名詞列に対してステップ 1 を適用した結果の方が，本手法より適合率が高い． この適合率に大きな差はないが，本手法で獲得された概念具体物関係の数は括弧つき名詞列から獲得された関係の 28%しかない．しかしながら，括弧つき名詞列から獲得された概念具体物関係と本手法で獲得された関係に共通する関係は存在せず，本手法を用いることで括弧つき名詞列から獲得できない概念具体物係が得られることが分かった．従って，近い将来，実用的な応用分野で必要となる概念具体物関係の数が不足した場合に本手法は有効であろうと考えられる．

最後に，日本語を対象とした上位下位関係の獲得における既存研究 [7, 6] で利用されている，他の語彙統語パターン (図 3) を用いて実験を行った． 4.4GB の HTML 文書集合にパターンを適用した結果，147,056 個の上位下位関係を獲得した．(既存研究中では構文解析済みの結果にパターンを適用しているが，本実験では単純な文字列のパターンによって上位下位関係を獲得した．) 獲得した上位下位関係からランダムに 200 個の関係を抽出し評価した結果，その適合率は 14.5%であった．適合率が低い原因は，最適化を行っていないためだと考えられる．仮に，本手法の評価で用いた 6.5GB の HTML 文書集合に対して図 3 に示したパターンを適用したとすれば，正しい上位下位関係は約 31,500 個，各パターンごとに平均して 3,937 個程度の正しい上位下位関係が得られると推測される．本手法は，これらのパターン一つあたりによって得られる上位下位関係の数と同程度の数の概念具体物関係を，明確な手がかりなしに高い適合率 83.5 %で獲得できることになる．

4 まとめ

本稿では，コーパス中に出現する単なる名詞列から概念具体物関係を抽出する方法について述べた．本手法

表 2: 括弧つき名詞列から得られた概念具体物関係

ステップ	概念具体物関係の数	適合率 [%]	予測される正しい関係の数
概念名制限前	38,987	55	21,442
概念名制限後	15,950	89	14,195

NP など NP	NP 以外の NP
NP などの NP	NP という NP
NP に似た NP	NP と言う NP
NP のような NP	NP と呼ばれる NP

図 3: 上位下位関係獲得のための括弧以外のパターン

は，**概念名を示しやすい名詞列が存在し，このような名詞列が前方の部分名詞列と一致する名詞列は概念具体物関係を示すことが多い**，という我々の直観に基づいている．更に，本手法では誤って抽出された概念具体物関係を取り除くために検索エンジンをフィルタとして用いた．本手法を用いることで，6.5GB の HTML 文書集合から 4,276 個の概念具体物関係が獲得され，その適合率は 83.5%であった．

今後は，より巨大な WWW 文書集合に本手法を適用し，獲得した概念具体物関係を情報検索や情報抽出のような応用分野で利用することを計画している．また，複合名詞解析 [8, 9] の技術を本手法に導入し，発展させることも今後の課題である．

参考文献

[1] Marti A. Hearst, “Automatic acquistition of hyponyms from large text corpora”, In Proceedings of the 14th International Conference on Computational Liguistics,pp.131–151, 1992.

[2] Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi, “Offline strategies for online question answering: Answering questions before they are asked”, In Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics,pp.1–7, 2003.

[3] Emmanuel Morin and Christian Jacquemin, “Automatic acquistion and expansion of hypernym links”, Computers and the Humanities, 38(4), pp.363–396, 2004.

[4] Sharon A. Caraballo, “Automatic construction of a hypernym-labeled noun hierarchy from text”, In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp.120–126, 1999.

[5] Deepak Ravichandran and Eduard Hovy, Towards terascale knowledge acquistion”, In Proceedings of the 20th International Conference on Computational Linguistics, pp.321–328, 2004.

[6] 安藤まや, 関根聡, 石崎俊. “定型表現を利用した新聞記事からの下位概念単語の自動抽出”, 情報処理学会 研究報告 2003-NL-157,pp.77–82, 2003.

[7] 今角恭祐, “並列名詞句と同格表現に着目した上位下位関係の自動獲得”, 修士論文, 九州工業大学, 2001.

[8] Mark Lauer, “Conceptuall association for compound noun analysis”, In Proceedings of the 32nd annual meeting on Association for Computational Linguistcs,pp.337–339 1994.

[9] Yoshiyuki Kobayasi, Takenobu Tokunaga, and Hozumi Tanaka, “Analysis of japanese compound nouns using collocational information”, In Proceedings of the 15th Conference on Computational Linguistics, pp.865–869 ,1994.