具象的な複合名詞における構成要素間の意味的な関係の分類

新里 圭司　　関根 聡
楽天技術研究所
{keiji.shinzato, satoshi.b.sekine}@.mail.rakuten.com

1　はじめに

本稿では「電動自転車」のような具象的な複合名詞において，修飾語 (e.g., 電動) と被修飾語 (e.g., 自転車) の間に成り立つ意味的な関係を明らかにし，その分類方法を提案する．このような意味的な関係は，英語では Nastase ら [7] や Moldovan ら [6] によって，約 30 種類の関係が提案されており，これらから選びだした *Cause-Effect* や *Part-Whole* など 7 種類の関係へ自動分類することを目的としたワークショップも開かれている [2]．一方日本語においては，「機械**操作**」のように被修飾語がサ変名詞となっている複合名詞の意味解析を試みた研究 [10] はあるが，要素間に成立する意味的関係を調査した研究はない．

我々の事前の調査によれば，楽天市場に入力されるクエリは「自転車」や「ヘルメット」のような具象名詞が多いことが分かっている．このようなクエリで検索すると，例えば「自転車」であれば「電動自転車」「子供用自転車」「自転車用ヘルメット」など様々な商品が混ざった検索結果が返される．これは検索されたページがクエリとして与えられた具象名詞を含んでいるためであり，この問題は楽天市場の検索に限らず，通常の情報検索システムにおいても起こる．本稿で述べる複合名詞に関する関係を自動獲得できれば，「目的」「ユーザ」等の観点から検索結果を分類できるようになり，ユーザの利便性を高めることが期待できる．

2　関連研究

複合名詞を構成する要素間の意味的関係を分類した研究は多くある [5, 8, 7, 6]．Levi[5] は修飾語と被修飾語の間に挿入されやすい前置詞を調査し，複合名詞を分類している．Nastase ら [7]，Moldovan ら [6] は実テキストを調べ，それぞれ 30 種類，35 種類の意味的関係を提案している．Moldovan らの分類は階層的でないのに対し，Nastase らの分類は階層的になっている．また Rosario ら [8] は医療ドメインのテキストに含まれる複合名詞を分類している．意味的関係を高精度で分類する手法としては Girju[3] の手法がある．Girju は Moldovan らの意味的関係から 22 種類選び，英語だけでなく，フランス語，イタリア語，スペイン語など 6 言語に対して，複合名詞の意味的関係を付与したデータを 6,200 件作成した．そして，複数言語の語彙統語パターンを素性とし分類器を学習することで 71.25%の精度を達成している．このように，近年は英語だけでなく，フランス語やイタリア語などの言語に対しても複合名詞の意味解析の研究が行われている．

表 1: 抽出された具象的な複合名詞の例.

出現頻度	文書頻度	店舗頻度	複合名詞
183,296	165,792	5,391	合成**皮革**
128,047	117,195	5,111	カラー**リング**
:	:	:	:
385	316	104	味付**数の子**
384	333	104	オーバー**キャップ**
:	:	:	:
10	10	10	オイル**スクリーン**
10	10	10	はと**麦粉**

一方日本語においては，竹内らによって語彙概念構造を用いて複合名詞の意味解釈を行う手法が提案されている [10]．これは「機械**操作**」や「機械**翻訳**」のように被修飾語がサ変名詞の場合を対象としたものであり，我々が今回対象とした具象的な複合名詞に対しては適用できない．また，複合名詞ではないが名詞句「A の B」の意味解析を試みた研究として Kurohashi ら，Torisawa の研究がある [4, 9]．Kurohashi らは国語辞典の定義文から，Torisawa は大量のテキストデータから名詞句の解釈として妥当な表現[1]を自動獲得するが，どちらの研究も得られた表現を「目的」や「ユーザ」等の観点からまとめあげることはしていない．

3　複合名詞の分類

どのような意味的な関係を用意すれば良いのかという問題は自明でない．そこで，実テキストより複合名詞を抽出し，複合名詞を構成する要素間にどのような関係が成り立っているのか調査した．

[1]例えば，「着物の女性」に対する「着物を着た女性」や「着物をまとう女性」.

表 2: 意味的な関係一覧 (M は修飾部，H は主辞を表す)

ID	関係名（事例数）	説明	語彙統語パターン	例（太字部分が主辞）
1	材料 (111)	H が M からできていることを表す.	M の H,M 製 H,M 入り H,M 製の H,M 入りの H,M 素材の H,M が入った H,M でできた H,M の入った H,M 地でできた H,M が配合された H	ファーベスト, 帆布トートバッグ, 皮手袋, トタン湯たんぽ, リネンストール, ニットワンピース, ナイロンジャケット, ナイロンベスト, シルバーリング, シルクストール, レザージャケット
2	目的 (77)	M は H の利用目的を表す.	M の H,M に H,M で H,M 用の H,M になる H,M に用いる H,M のための H,M 性の高い H,M 計測用の H,M の効果がある H,M するために使う H,M 効果が得られる H,M 効果の得られる H	ウォーキング用ソックス, ゴルフマーカー, 学習デスク, 水泳キャップ, 安全靴, 携帯用スリッパ, 浴用石鹸, バスケシューズ, ラップタイマー, 発熱レッグウォーマー, エコブランケット
3	サイズ (43)	M は H のサイズを表す.	M な H,M の H,M 丈の H,M サイズ H,M サイズの H	ロング手袋, ミニドレス, コンパクト双眼鏡, ハーフパンツ, ミニバッグ, シングルソファー, 大型トートバッグ, ハンディテーブル, 長袖トレーナー, ミニ手帳, ショートブーツ
4	対象物 (42)	M のために H を使うことを表す.	M 用 H,M 用の H,M で使う H,M に巻く H,M に付ける H,M に用いる H,M のための H,M を付ける H,M を入れる H,M テープ用の H,M 品をいれる H,M を入れるための H	フィルターケース, エンジンカバー, リネンスプレー, パフケース, フェイスシート, ズボンハンガー, ジュエルケース, 冷茶ポット, 眼鏡ケース, ＯＡラック, ストロベリーポット
5	形 (41)	M は H の形を表す.	M の H,M 型の H,M 状の H,M タイプの H,M の形をした H	棒ネクタイ, ラウンドファスナー財布, 馬蹄クッション, 円座クッション, ペタンコ靴, バルーンワンピース, ノースリーブニット, エプロンドレス, ペンダントネックレス, フックピアス, アニマル湯たんぽ
6	ユーザ (33)	M は H の利用主体を表す.	M の H,M 用 H,M 用の H,M が使う H,M ものの H,M 向きの H,M 向けの H	メンズ福袋, ベビーリュック, メンズ財布, レディシェーバー, レディース腕時計, 幼児用自転車, 子供用スーツケース, 男性用腕時計, メンズスニーカー, 子供用リュック, 子供靴
7	付属品 (27)	M に H が付属していることを表す.	M の H,M 付き H,M 付きの H,M が付いた H,M の付いた H	袷着物, クロスピアス, チェーンネックレス, ダイヤモンドピアス, スパンコールネクタイ, ファーマフラー, レースストール, チェーンショルダーバッグ, モップスリッパ, フード付きカーディガン, レースカーディガン
8	雰囲気 (26)	M は H が持っている雰囲気を表す.	M な H,M の H,M 調の H,M 風の H,M のような H,M の際に着る H,M が使うような H,M が着ているような H	家具調こたつ, ギャルソンエプロン, 和風こたつ, 姫ドレス, サンタ帽子, フライトジャケット, メッセンジャーバッグ, ボレロ風カーディガン, おしゃれ着物, 和帽子, オリジナルランドセル
9	機能 (25)	M は H が備えている機能を表す.	M の H,M 内蔵 H,M 型の H,M 式の H,M で動く H,M 機能付き H,M 機能がついた H	ＰＴＣヒーター, 防寒ズボン, 防水コンセント, コードレスヘッドホン, ＢＬＵＥＴＯＯＴＨブレスレット, マルチポーチ, フローティングベスト, 耐熱カップ, 防水携帯, 折りたたみ式自転車, 保冷ポット
10	場所 (18)	M は H が利用される場所を表す.	M の H,M 用の H,M で使う H,M で使われる H,M で使用する H,M で利用される H,M にいる選手が用いる H	ガーデンエプロン, リビングこたつ, カフェエプロン, 楽屋ミラー, スクールネクタイ, サロンエプロン, ショップクリーナー, センターキャップ, トイレスリッパ, ルームスリッパ, 水中ヒーター
11	柄 (17)	M は H に施された柄を表す.	M の H,M 柄の H,M された H	ボーダーワンピース, チェックストール, チェック柄ワンピース, ボーダーカーディガン, 千支タオル, チェックネクタイ, ストライプネクタイ, 花柄ワンピース, ボーダーマフラー, 花柄トートバッグ, チェック柄ストール
12	色 (16)	M は H の色を表す.	M な H,M の H,M 色の H,M カラーの H	ホワイトダイヤモンド, ピンクダイヤモンド, カラーパンツ, ブラックダイヤモンド, 黒ネクタイ, グリーンカレー, イエローダイヤモンド, ブルーダイヤモンド, カラーブロック, ブラウンダイヤモンド, レッドカレー
13	加工 (14)	M は H に施された加工を表す.	M が施された H,M 加工された H	シールドケーブル, メッシュジャケット, デザインリング, エナメルリュック, シールドバッテリー, 裏毛トレーナー, メッシュベルト, エナメルスニーカー, エナメルショルダーバッグ, ホログラムシート, デザインネックレス
14	使用法 (9)	M は H の使用法を表す.	M で使う H,M して使う H,M として使う H	斜め掛けショルダーバッグ, ペアリング, 斜めがけショルダーバッグ, 手持ち花火, 背もたれクッション, インナー手袋, ペアネックレス, クラッチバッグ, ラップスカート
15	仕様 (8)	M は H の仕様を表す.	M 型の H,M 的な H,M できる H,M 仕様の H,M 形式の H	使い捨て手袋, ワンタッチボトル, アナログ腕時計, 電子手帳, ロッキンググラス, ワンタッチクリップ, デジタル腕時計, 四つ身着物
16	性質 (8)	M は H が備えている性質を表す.	M な H,M の H,M 型の H,M 性の H,M の性質をもった H	薄手ニット, ナチュラルシャンプー, ソフトクッション, ハードスーツケース, 軽量スーツケース, レトルトカレー, ストレッチパンツ, 弱酸性シャンプー
17	状態 (6)	M は H の状態を表す.	M の H,M 状の H,M された H	天然ダイヤモンド, スープカレー, リサイクル着物, 液体のり, オーガニック野菜, 中古タイヤ
18	産地 (6)	M は H の生産された場所を表す.	M 産 H,M 産の H,M で使っている H	タイカレー, ダージリンティー, アフガンストール, トルココーヒー, 汐吹昆布, インドカレー
19	着用箇所 (6)	M は H を着用する場所を表す.	M 用の H,M に付ける H,M からかける H,M からさげる H,M からかけて使う H	ウエストポーチ, ショルダーベルト, ネックポーチ, ウエストバッグ, ウエストベルト, ショルダーポーチ
20	コンテンツ (5)	M は H の中身を表す.	M が入った H,M の品が入った H,M を書くための H	テーピングバッグ, ブランド福袋, スイーツ福袋, アクセサリー福袋, スケジュール手帳
21	時期 (4)	M は H を利用する時間を表す.	M の H,M 用 H,M 用の H,M の時期の H	クリスマス福袋, サマードレス, 新春福袋, サマーニット
22	駆動 (4)	M は H の駆動方式を表す.	M の H,M 式の H,M 形式の H	ウォーターオーブン, 電気こたつ, 機械式腕時計, ＵＳＢブランケット
23	生産者 (4)	M は H を生産した主体を表す.	M の H,M 製の H	グッチネクタイ, デザイナーズエプロン, ブランドネクタイ, ＮＧＫプラグ
24	順序 (3)	M は H の順序を表す.	M 的な H	セミショルダーバッグ, サブバッグ, サブリュック
25	評価 (2)	M は H の評価を表す.	M のある H	人気スニーカー, 人気トートバッグ
26	価格 (2)	M は H の価格を表す.	M の H	激安福袋, アウトレット福袋
27	方式 (2)	M は H の方式を表す.	M の H,M 方式の H	プレス式コーヒーメーカー, サイフォン式コーヒーメーカー

3.1 具象的な複合名詞の獲得

本研究では簡単のため，以下のパターンに適合する表現を具象的な複合名詞とし，分類の対象とする．

名詞 **<具象名詞>** (e.g.,電動 **自転車**)

名詞 接尾辞 **<具象名詞>** (e.g.,子供用 **自転車**)

以下では，下線部を**修飾部**，太字部分を**主辞**と呼ぶ．<具象名詞>としては，分類語彙表[2]の「体:自然:物質」「体:自然:動物」「体:自然:植物」「体:生産物 (土地利用は除く)」カテゴリ以下に属す 13,432 表現を用いた[3]．

続いて，楽天データ公開[4]より提供されている楽天市場の商品データ（商品名および商品説明文）に対して上記のパターンを適用した．この時，形態素解析器には MeCab[5]を用い，品詞が未知語および記号-アルファベットとなった形態素は名詞として扱った．反対に「お茶」や「1 袋」等を除くため，品詞が名詞-接尾，名詞-数となっている形態素は名詞と見なさなかった．さらに，分類語彙表に登録されている表現は 1 語と見なし除いた．

上記のパターンでは「らリラ」や「だんご剣」のような表現も獲得されてしまう．そこで，このような表現を除くため，2011 年 6 月に 10 回以上検索キーワードとして入力された表現のみに絞った．

最終的に 27,160 個の複合名詞が得られた．抽出された複合名詞の例を表 1 に示す．店舗頻度とは，その複合名詞を用いた店舗の異なり数を表す．以降，この複合名詞の集合を**複合名詞プール**と呼ぶ．

3.2 意味的関係の調査

以下の 2 種類のセットを対象に，修飾部-主辞間に成り立っている意味的関係を調査した．

セット A: 検索回数の高い上位 50 件の具象名詞 (e.g., 靴) について，それぞれが主辞に現れている複合名詞 (e.g., 革**靴**やスキー**靴**) を店舗頻度の高い順に 10 件ずつ複合名詞プールより抽出して得た 448 件の複合名詞集合

セット B: 複合名詞プールから無作為に抽出した 200 件の複合名詞集合

複合名詞にも多義性の問題がある．例えば「リボン**バスケット**」という表現はリボンの付いたバスケットとリボン柄のバスケットという 2 つの解釈が考えられる．単純に複合名詞を見ただけではどちらの語義か判断できない．そこで，楽天市場の商品データから複合名詞を含むものを無作為に 15 件抽出し，最も頻出する語義をその複合名詞の語義とした．商品データを抽出する際は店舗が被らないよう注意した．複合名詞の意味がわからない場合はウェブを調べ，分類と同時に修飾部–主辞間に成り立つ語彙統語パターンを列挙した．

調査の結果，複合名詞 648 件のうち 580 件に意味的関係があり，残りの 24 件は固有名詞もしくは一語とみなした方が良い表現，44 件は解析誤りによって単名詞が過分割された表現や主辞が具象名詞でない表現であった．得られた意味的関係一覧を表 2 に示す．1 事例しかない関係はその複合名詞に特化したものと判断し除いている．

4 実験

4.1 分類基準の評価

3.2 節の調査に用いなかった複合名詞を，複合名詞プールより無作為に 200 件抽出し，表 2 にあげた関係への分類を行った．被験者は著者の 1 人及び著者以外の人物 1 人である．

先述したように，複合名詞にも多義性の問題がある．そこで，3.2 節同様，複合名詞を含む商品データ 15 件を被験者に文脈として提示し，最も頻出する語義をその複合名詞の語義と見なすよう指示した．

分類の際は，表 2 を提示し，関係の説明，語彙統語パターン，事例を参考にして分類するように伝え，複数個の関係が考えられる場合は，全ての関係に分類してもらった．また，与えた文脈だけからでは関係の判断が難しい場合は，ウェブを調べるよう指示した．

実験の結果，被験者間の κ 統計量は 0.681 であった．これは Substantial 程度の一致を意味している．被験者数が少ないが，この結果は意味的関係の説明，語彙統語パターン，事例の提示が分類の基準として妥当であることを示唆している．

いずれかの被験者がその他（表 2 で与えた意味関係以外）に分類した複合名詞は 15 件，被験者が共に解析誤りと判定した複合名詞は 24 件であった．この結果から 91.5 $(=100\times(161/176))$%の複合名詞を表 2 の関係でカバーできたことがわかる．

4.2 複合名詞の自動分類

事例数が 10 以上ある意味関係について自動分類を試みた．3.2 節で調査に用いた 668 件の複合名詞を 5 分割し，交差検定を行った．分類に用いた素性を表 3 に

[2]http://www.ninjal.ac.jp/products-k/kanko/goihyo/
[3]ただし,「商品」「本体」「カラー」「素材」「オリジナル」「幅」「タイプ」は一般的すぎるため用いない．
[4]http://rit.rakuten.co.jp/rdr/index.html
[5]http://mecab.sourceforge.net/

表 3: 利用した素性

素性名	素性数	説明
語彙統語パターン	99	表 2 に示した語彙統語パターンに修飾部，主辞を当てはめた表現がウェブ上に存在するか否か †
修飾部	595	修飾部の文字列
品詞	7	修飾部の品詞（第 1 層と第 2 層）
接尾辞	19	修飾部末尾にある接尾辞
意味クラス	55	修飾部が属す分類語彙表の意味クラス（上位 3 階層）

†: ヒット件数の確認には Yahoo! Web API を利用

表 4: 自動分類結果

関係名	精度	再現率	F 値	関係名	精度	再現率	F 値
材料	0.816	0.788	0.802	雰囲気	0.413	0.328	0.366
目的	0.708	0.590	0.644	機能	0.440	0.350	0.390
サイズ	0.797	0.920	0.854	場所	0.400	0.450	0.424
対象物	0.542	0.495	0.517	柄	0.817	0.691	0.748
形	0.639	0.544	0.587	色	0.497	0.686	0.576
ユーザ	0.880	0.814	0.846	加工	0.450	0.450	0.450
付属品	0.538	0.571	0.554	**Ave.**	**0.611**	**0.590**	**0.600**

示す．学習器には MIRA[1] の C++実装を用いた[6]．

実験結果を表 4 に示す．Girju[3] とは実験設定が異なるため単純な比較はできないが，Girju よりも学習に用いたデータ数が少ないことを考慮すれば，表 4 に示した精度は大きく劣っていないと考えられる．次に分類結果の混同行列を表 5 に示す.「コケ取り**スプレー**」や「エコ**ブランケット**」など「目的」に分類されるべき表現を「その他」に分類してしまう誤りが最も多かった．その多くは,「M の H」や「M な H」などの語彙統語パターン素性しか発火していなかった．これらは他の意味関係でも用いられる曖昧なパターンであるため誤ったと考えられる．

5 おわりに

本稿では具象的な複合名詞（e.g., 電動自転車）において，修飾語（電動）と被修飾語（自転車）の間にどのような意味的な関係があるのかを実テキスト通して調査し,「目的」「対象物」「ユーザ」など 27 種類の関係を定義した．そして，各関係に対して人手で与えた語彙統語パターン等を素性とする分類器により 60%の精度で分類できることを示した．

今後の課題としては分類精度の向上が挙げられる．現在は語彙統語パターンを人手で与えているため網羅性にかける．そのため，事例をシードとしたパターンの自動獲得によりその網羅性を高めることで精度の向上が期待できる．また，2 名の被験者で分類を行い高い一致率を示したが，被験者数を増やしても同様の結果が得られるかどうか調査することも今後の課題である．さらに自動分類された複合名詞を利用したファセット検索への応用についても今後考えたい．

[6]http://www.r.dl.itc.u-tokyo.ac.jp/study_ml/pukiwiki/index.php?manual からダウンロードできる．

表 5: 自動分類結果の混同行列

		システムの出力														
		材料	目的	サイズ	対象物	形	ユーザ	付属品	雰囲気	機能	場所	柄	色	加工	その他
正解データ	材料	**89**		1	3	1		4	1	1				2	9
	目的		**40**	1	6			1	1	4	3				11
	サイズ			**39**			2								2
	対象物	1	5	1	**21**	1	1	1			1	1			9
	形	2	1	2	1	**21**		5	2		1				3
	ユーザ				1		**28**		1						3
	付属品	4			1	5		**15**							2
	雰囲気		1			3			**9**	1	1				9
	機能		3	1	1	2				**10**				2	5
	場所	1	2			1					**6**		1		7
	柄				1			1	1			**11**	1		2
	色	1											**11**		4
	加工	3				1		1		1				**5**	3
	その他	9	6	5	5	7	2	4	4	2	6	2	4	2	**103**

参考文献

[1] Koby Crammer and Yoram Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991, 2003.

[2] R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Semantic Evaluation Workshop (SemEval)*, 2007.

[3] Roxana Girju. The syntax and semantics of prepositions in the task of automatic interpretation of nominal phrases and compounds: a cross-linguistic study. *In Computational Linguistics*, 35(2), 2009.

[4] Sadao Kurohashi, Masaki Murata, Yasunori Yata, Mitsunobu Shimada, and Makoto Nagao. Construction of japanese nominal semantic dictionary using “a no b” phrases in corpora. In *Proceedings of COLING-ACL’98 Workshop, The Computational Treatment of Nominals*, 1998.

[5] Judith N. Levi. *The Syntax and Semantics of Complex Nominals*. Academic Press, New York, 1978.

[6] Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. Models for the semantic classification of noun phrases. In *In HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 60–67, 2004.

[7] Vivi Nastase and Stan Szpakowicz. Exploring noun-modifier semantic relations. In *Proceedings of the 5th International Workshop on Computational Semantics*, 2003.

[8] Barbara Rosario and Marti Hearst. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01*, pages 82–90, 2001.

[9] Kentaro Torisawa. A nearly unsupervised learning method for automatic paraphrasing of japanese noun phrases. In *Proceedings of the Workshop on Automatic Paraphrasing*, 2001.

[10] 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**, 43(5):1446–1456, 2002-05-15.