形容詞を用いた対象・属性名詞対の収集および分析

阿辺川 武
東京工業大学大学院 総合理工学研究科
abekawa@lr.pi.titech.ac.jp

奥村 学
東京工業大学 精密工学研究所
oku@pi.titech.ac.jp

1 はじめに

近年，Web 上の情報を利用するための技術として情報抽出への期待が高まっている．特に，ある対象に対する評判情報の抽出はマーケティングや企業のリスク管理の点から広く求められている．評判情報は主に対象，属性，評価値の 3 つ組から成り立っており，それらの 3 つ組の辞書を作成する必要がある．従来では人手により作成されていたが，人手による辞書の作成では，網羅性や作成のコスト問題が生じる．そこで Web コーパスから評価表現を自動収集する研究が数多く提案されている [3, 6]．ただし，これらの手法はドメイン依存であったり，膨大な 3 つ組データが必要になる．

また，情報検索において膨大な検索結果から必要な情報を選択あるいは統合するために，ある対象に対して関連する属性を抽出する研究が試みられており，Web の表形式から属性表現を収集する手法 [5] や，ルールを用いて属性表現を収集し，オトロジーを構築する手法などが提案されている [4]．

本論文では,「鼻が長い象」や「料理がおいしいレストラン」などのような「名詞 A+が+形容詞+名詞 B」という構文において,「名詞 A が」が形容詞に係るとき，名詞 A と名詞 B が属性と対象の関係になっていることが多いという前提のもとに，Web 上からそのような表現を収集し，ある対象名詞について，その属性集合を収集する手法を提案する．

本稿では最初に，ある対象名詞について Web 上から上記構文を収集し，人手により「名詞 A が」が形容詞に係る事例のラベル付けを行う．そして，そのような事例から名詞 A を抽出し，実際に属性名詞となっているかを確認する．次に人手によりラベル付けされたデータから「名詞 A が」が形容詞に係るか否かを識別する学習モデルを構築し，別の対象名詞について機械学習アルゴリズムを用いた属性名詞抽出を試みる．最後に得られた属性名詞集合についての考察を行なう．

2 対象名詞・属性名詞のペアの収集

本提案手法では「名詞 A+が+形容詞+名詞 B」という構文において,「名詞 A が」が形容詞に係るとき，名詞 A と名詞 B が属性と対象の関係になっていることが多いという前提を利用する．

「名詞 A+が+形容詞+名詞 B」という構文に着目した理由は，形容詞はその多くが必須格を 1 つしか持っていないからである.「名詞 A が」が形容詞に係り，形容詞の格スロットを埋めているとき，名詞 B は形容詞に連体修飾されていても形容詞との格関係を持つことが出来ず，名詞 A と関係を持つことになるからである．

2.1 収集方法

本節では対象名詞「レストラン」を例に，Web 上のコーパスからその属性名詞集合を収集する手法を説明する．

2.1.1 フレーズ検索

ある対象名詞に対する上記の構文を検索エンジンを用いて収集する．本研究では，対象名詞が上記の構文で使用されている事例をできるだけ多く収集したい．そこで，新聞コーパスで頻出するイ形容詞，ナ形容詞をそれぞれ 500 個ずつ用意し，それぞれの形容詞に対して “が形容詞レストラン” といったフレーズ検索を行い，そのフレーズを含む snippet を収集する．検索エンジンには Yahoo! Japan[1]を使用した．

2.1.2 フィルタリング

フレーズ検索では，間に記号が挿入されていても，それらは無視される．例えば “がおいしいレストラン” でフレーズ検索を行っても「〜がおいしい。レストランでは〜」のように間に句点が入ったページも検索されてしまう．そのような snippet を削除するために，文字列マッチングを用いて，間に記号が挿入されているものは削除する．さらに「名詞 A+が+形容詞+名詞 B の名詞 C」の場合では形容詞の係り先に曖昧性が生じてしまうので，名詞 B に助詞「の」が後接する場合も同様に削除する．その後，句点などの記号を文区切りとみなし，snippet からフレーズが含まれる文を抽出する．

検索エンジンは内容が全く同じページでもホストが異なれば，別々の検索結果のリストとして表示されることがあるため，文字列として完全に同じ文は 1 文に

[1]Yahoo!検索 Web サービス
http://developer.yahoo.co.jp/search/

まとめている．最終的には，取得 snippet 数が 5,545 個に対し，フィルタリング後の文数は 2,512 文となった．

2.2 構文の分類

得られた文集合を，名詞 A の係り先と対象・属性の観点から以下の 4 つに分類する．

ラベル 1 名詞 A+が+形容詞に係り，名詞 A が名詞 B の属性である．
「予約が困難なレストラン」
「ローカル色が強いレストラン」

ラベル 2 名詞 A が名詞 B 以降の文節に係る．
「彼氏が美味しいレストランに連れて行って〜」
「ここが高級なレストランである」

ラベル 3 「名詞 A が」が形容詞に係るが，名詞 A が名詞 B の属性ではない．
「私が好きなレストラン」
「ペット同伴が可能なレストラン」

ラベル 4 文区切りの失敗のため得られた文字列が文を成していない．対象名詞が文節の主辞となっていない．
「一人がおしゃれなレストラン兼雑貨ショップ見」
「読者コメントが面白いレストラン情報」

ラベル付与は著者が 1 人で行った．分類の内訳を表 1 に示す．

表 1: 分類の内訳

ラベル 1	1,715
ラベル 2	325
ラベル 3	433
ラベル 4	39

ラベル 1 に分類した事例から名詞 A および名詞 A+形容詞のペアを抜き出し，頻度順に列挙したものが表 2 である．ただし，検索エンジンの制約上，1 つのフレーズ検索について得られる snippet の上限が 1000 個であり，また同一の文を 1 文にまとめているため，実際の Web 上の頻度とは大きく異なることに注意していただきたい．

名詞 A のリストを見ると，ほとんどの名詞は対象名詞「レストラン」の属性名詞であると考えられる．これにより「名詞 A+が+形容詞+名詞 B」の構文を用いることで Web 上から対象名詞と属性名詞のペアが収集できることが確認できた．

得られた属性名詞リストには，料理や食べ物といった属性の他にも「夜景」「眺め」「眺望」などの景色に関する属性,「雰囲気」「居心地」といった居住性に関する属性,「内装」「インテリア」「外観」のような建物に関する属性など，多様な種類の属性が存在している．

3 属性名詞集合の自動収集

前節では「名詞 A+が+形容詞+名詞 B」の構文により属性名詞集合が収集できるかについて，実際に人手によりラベル付けを行ない確認した．本節では，ラベル付けして得られた事例を学習データとみなして，機械学習により「名詞 A が」が形容詞に係るか否かを識別するモデルを学習する．そしてある対象名詞に対してその属性名詞集合を自動的に収集することを試みる．

3.1 素性について

機械学習アルゴリズムで使用する素性について説明する．例文としてラベル 1 である「小娘には敷居が高

表 2: 収集した属性名詞の例

名詞 A		名詞 A+形容詞	
122	夜景	72	夜景 綺麗だ
89	料理	43	料理 美味しい
81	雰囲気	36	予約 必要だ
65	予約	29	パン 美味しい
63	眺め	27	人気 高い
46	景色	23	夜景 美しい
38	パン	22	予約 困難だ
28	眺望	22	雰囲気 良い
28	人気	21	雰囲気 いい
25	インテリア	19	眺め いい
20	内装	18	夜景 素敵だ
17	居心地	14	雰囲気 素敵だ
16	評判	14	景色 いい
16	ビール	13	眺め 良い
15	種類	11	敷居 高い
14	値段	11	眺望 素晴らしい
13	窓	11	値段 高い
13	気持ち	11	種類 豊富だ
13	シーフード	10	料理 有名だ
12	敷居	10	料理 豊富だ
12	ピザ	10	雰囲気 よい
12	サービス	10	眺め 美しい
11	展望	10	眺め 素晴らしい
11	天井	10	インテリア 印象的だ
11	建物	9	景色 良い
11	外観	8	評判 良い
10	ワイン	8	ピザ 美味しい
9	料金	7	夜景 素晴らしい
9	見晴らし	7	天井 高い
9	魚介類	7	景色 素晴らしい
9	飲茶	7	居心地 良い
8	夕日	7	気持ち いい
8	評価	7	管理 必要だ

いレストランだが、高級すぎるほどではない」を用いる．最初に文を形態素解析した後，表 3 のように構文を含む文節とその 1 つ前の文節を取り出す．そして，それぞれの文節について表 4 にある素性を抽出する．なお形態素解析器には JUMAN を，文節区切りには KNP を利用した．

表 3: 素性を抽出する文節

	文節 1	文節 2 (名詞 A)	文節 3 (形容詞)	文節 4 (名詞 B)
例	小娘には	敷居が	高い	レストランだが,

表 4: 使用した素性

素性	例
1 文節 1 の語形の見出し	は
2 文節 1 の語形の品詞	助詞
3 文節 1 の語形の品詞細分類	副助詞
4 文節 1 の読点の有無	無
5 文節 2 の主辞の見出し (接尾辞がある場合は接尾辞も含める)	敷居
6 文節 2 の主辞の品詞	名詞
7 文節 2 の主辞の品詞細分類	普通名詞
8 文節 2 の主辞の活用	無
9 文節 2 の主辞の活用形	無
10 文節 2 の主辞の見出しの概念 (固有名詞は深さ 2，一般名詞は深さ 3)	0533
11 文節 3 の主辞の見出し	高い
12 文節 2 と文節 4 の概念が同一か	異なる
13 文節 4 が文中か文末か	文中

文節の語形・主辞の定義は，文献 [7] にならう．また語の概念は，日本語語彙大系 [2] を参照し，ルートからの一定の深さにある概念番号を使用する．

3.2 実験・結果

まず，2.2 節でタグ付けを行ったデータに対して，分類実験を行なった．実験に用いたデータは，ラベル 4 を除いたもので，ラベル 1,2,3 の 3 値分類となる．

機械学習アルゴリズムには SVM[2]を使用し，カーネルは線型カーネル，多値分類には pair wise 法を用いた．評価は 5 分割交差検定で行った．

実験の結果を表 5 に示す．正解率とは，分類器が正しくラベルを判別したときの正解率，ラベル 1 の精度とは，分類器がラベル 1 と判別した内，正解した事例の割合，ラベル 1 の再現率とは，全てのラベル 1 の事例に対して，分類器がラベル 1 と判別した割合である．

[2]SVM の実装には TinySVM http://chasen.org/~taku/software/TinySVM/ を用いた

表 5: SVM による「レストラン」の分類精度

正解率	0.945
ラベル 1 の精度	0.955
ラベル 1 の再現率	0.979

3.3 他の対象名詞における属性名詞集合の収集

得られた学習モデルを利用して他の名詞について，その名詞を対象名詞として属性名詞集合を収集する実験を行った．もちろん対象名詞「レストラン」において学習したモデルなので，語の見出しを素性とするものは「レストラン」に特化した素性のままである．

実験には，対象名詞として「ホテル」「会社」「デジカメ」を選択した．それぞれの名詞について，2 節の手法で Web から文を収集し，3.2 節で用いた学習モデルを使用して構文を分類した．収集された文数，ラベル 1 の文数，そしてランダムに選択した 200 個の事例に対して人手で評価した分類正解率を表 6 に掲載する．また，それぞれの対象名詞について得られた属性名詞集合のリストを表 7 に掲載する．

表 6: 他の対象名詞における実験結果

名詞	ホテル	会社	デジカメ
取得 snippet 数	12,411	26,075	4,629
フィルタリング後	5,090	10,348	2,203
ラベル 1 に分類	3,609	7,120	955
正解率 (任意 200 個)	95.5%	81.0%	95.0%

3.4 考察

3 つの名詞についてランダムに 200 個を抽出し評価した分類正解率では「ホテル」「デジカメ」は「レストラン」よりも良い正解率を示している．その理由として「ホテル」は「レストラン」とドメインが似通っているため，また「デジカメ」は名詞 A に「私が」「夫が」のように人称名詞が来る事例が多く，名詞 A に関する素性が有効に働いたためであると考えられる．

表 7 により得られた名詞 A のリストを見ると,「レストラン」と同様に，対象名詞の属性名詞となっている名詞が多い．しかし，対象名詞「会社」に対する「会社」のように，属性とは言えないものもある．対象名詞が異なる学習データを用いたことが主な原因であり，今後は対象名詞に依存しないモデルを構築する必要がある．

表 7 では頻度が上位の名詞を掲載しており，多くが属性名詞であると考えられる，低頻度の方の名詞では，分類誤りを含め，属性とは言えないものが含まれていた．分類結果から属性名詞集合を得ようとしたとき，

頻度を閾値として属性であるかを判断するか，それとも人手で判断するかは，現在考慮中である．

今回は属性名詞の候補となる名詞を 1 形態素 (接尾辞を含む場合は 2 形態素) としている．「(交通の) 便」「(資金) 繰り」「(電池の) 持ち」のように本来は複数形態素で意味を持つ語については，なんらかの基準を用いて複数語として扱う必要がある．

この手法は Web コーパスを利用し，対象名詞と共に高頻度で共起する名詞 A を属性として考えている．そのため対象名詞が高頻度で Web コーパスに出現する一般名詞ならばある程度大規模な属性名詞集合が得られるが，頻度が低い固有名詞では，大きな属性名詞集合が得られない可能性がある．そこで固有名詞については，固有名詞から上位概念である一般名詞を求め，両者を対象名詞として属性名詞集合を収集することを考えている．

4 おわりに

本論文では,「名詞 A+が+形容詞+名詞 B」という構文に着目し，対象名詞と属性名詞のペアを収集する手法を提案した．まず例として対象名詞「レストラン」について人手でラベル付けを行ない，実際に属性名詞が収集できることを実証した．そして「レストラン」のデータから得られた「名詞 A が」が形容詞に係るか否かを識別する学習モデルを使用して，他の対象名詞についても同様に属性名詞集合が取得できることを確認した．提案手法では「名詞 A+が+形容詞+名詞 B」において形容詞だけを扱ったが,「〜が自慢の〜」「〜が評判の〜」「〜が人気の〜」のような名詞を使用した表現についても，同様に対象名詞・属性名詞のペアが収集できると考えている．

今後の課題として，今回の手法で取得できる属性名詞が，対象名詞の持つ本来の属性のうちどのくらいの割合で収集できているのかを調査する必要がある.「名詞 A+が+形容詞+名詞 B」という構文だけで取得できない属性名詞があるならば，他の表現を利用しなければならない．

参考文献

[1] 橋本泰一, 白井清昭, 徳永健伸, 田中穂積. 統計的手法に基づく形容詞または形容動詞の修飾先の決定. 情報処理学会研究報告 138-NL-12, pp. 87–94, 2000.

[2] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系 – 全 5 巻 –. 岩波書店, 1997.

[3] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 3, pp. 203–222, 2006.

[4] 松平正樹, 上田俊夫, 大沼宏行, 森田幸伯. Web コンテンツの分析に基づくオントロジー構築および情報整理の試み. 人工知能学会，センマンティックウェブとオントロジー研究会,SIG-SWO-A302-08.

[5] 大前信弘, 黄瀬浩一. Web の表を対象とした属性の自動識別. 情報処理学会研究報告 171-NL-8, 2006.

[6] 鈴木泰裕, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. 人工知能学会，センマンティックウェブとオントロジー研究会,SIG-SWO-A401-02.

[7] 内元清貴, 関根聡, 井佐原均. 最大エントロピー法に基づくモデルを用いた日本語係り受け解析. 情報処理学会論文誌, Vol. 40, No. 9, pp. 3397–3407, 1999.

表 7: 取得した名詞 A のリスト

ホテル	会社	デジカメ
179 夜景	162 規模	67 画素数
138 アクセス	146 業績	44 機能
128 外観	115 会社	40 性能
79 便	89 関係	24 スピード
76 サービス	83 出入り	20 解像度
59 眺め	81 給料	19 画質
58 料金	78 経営	19 マクロ
54 部屋	71 対応	18 撮影
54 人気	66 比率	17 ズーム
49 値段	61 年齢	14 操作
46 雰囲気	59 意識	14 レンズ
40 料理	54 休み	13 反応
37 景色	53 入れ替わり	13 調子
36 眺望	53 審査	13 接写
36 インテリア	51 仕事	13 画素
35 建物	49 歴史	11 持ち
35 居心地	48 人数	11 サイズ
33 規模	48 条件	10 時間
32 壁	46 手数料	10 運び
32 風呂	46 管理	10 サイクル
30 立地	45 レベル	9 写真
28 感じ	41 評判	9 起動
25 評判	41 繰り	9 モード
25 パフォーマンス	40 株価	8 倍率
24 設備	38 時間	8 動画
24 屋根	36 知名度	8 速度
23 朝食	36 居心地	8 深度
23 対応	35 人使い	8 画像
22 条件	34 料金	7 感度
22 勝手	32 雰囲気	7 カメラ
21 プール	32 懐	6 評価
21 ビーチ	31 離職率	6 値段
20 ロケーション	30 内容	6 勝手