大規模テキストからの複合語の属性表現の抽出手法

臼渕 護　平手 勇宇　山名 早人
早稲田大学大学院理工学研究科
{usubuchi,hirate,yamana}@yama.info.waseda.ac.jp

1　はじめに

近年 Web 上に大量に存在する掲示板やブログ等のテキストデータから，評判情報を収集，解析する技術に注目が集まっている．テキストデータに含まれる評判情報の抽出には，ドメインごとに評価対象の性質や一部分を表す属性表現の辞書を作ることが重要である．この際，属性が「高画質モード」のように複合語からなる場合，「モード」という一段階上の概念を表す上位語として扱うか，そのまま複合語として扱うかが問題であった．

従来研究[1][2]では両方の手法が提案されており，これらは精度向上のため半手動で辞書作成が行われている．複合語を上位語として扱った場合の辞書構築は登録すべき表現が少ないため，辞書構築量という観点からは容易である．しかし，例えばプリンタドメインにおいて「フォトモード」と「高画質モード」等の区別ができないことに加え，「i モード」などのドメインに関連性のない語句を属性表現として扱ってしまうなど，精密さに欠ける．これに対し，複合語をそのまま扱った場合，辞書の精度を上げることができ，より詳細に評判情報を扱えるが，作成時に多くの語句を登録する必要があるため，多大なコストがかかる．

本論文ではこうした問題に対し，まず存在するすべての複合語に関して検索エンジンを利用してドメイン関連度を求め，閾値以下のものをフィルタリングする処理を行う．こうすることでドメインに関連性のない語句を排除できるため，上位語が等しい語句同士は意味的に大きく異なることがなくなり，同一語句して扱うことができると考える．つまり前述の例の場合，「フォトモード」と「高画質モード」はプリンタの印刷形態を表す「モード」の一種であるというとらえ方をすれば，辞書作成の際に同じ語句として扱っても問題ないと考える．次にこれらの語句を同一視して辞書の作成を行うことにより，一つ一つ別の語句として扱う必要がなくなり，低コストで網羅性が高く，かつ精度の高い辞書の作成を実現することが可能となる．

2　既存研究との比較

小林ら[1]は，文書に含まれる意見が *<対象，属性，評価>* の 3 要素からなると考え，属性表現を評価表現と共に，独自に作成した文型パタンを利用して，半手動でブートストラップ的に抽出する手法を提案した．この手法では，例えばプリンタドメインにおいて「高画質モード」等の複合語を「モード」という最小の単位で扱っており，複合語のように単語の組み合わせを考えなくてよいため，辞書に登録すべき表現の数が少なくて済み，短時間で網羅的な収集が可能である．しかし「i モード」のようなドメインに無関係の属性表現とはいえない語句との区別ができない．また「デジタル画像」と「アナログ画像」の評価の比較などができず，抽出される評判情報は大雑把なものになる．

そこでより精密な辞書を作るため小林ら[2]は従来手法に複合語も抽出対象とできるような処理を加え，抽出を行った．その結果，大量の属性表現を獲得することができるようになり，辞書の精度は上がり，より詳細に評判情報を扱えるようになった．しかし，個々の出現頻度が少なくなったことに加え，抽出を半手動で行っているため，多大な労力とコストがかかるという問題を持つ．特にドメインごとに辞書を作成したい場合には，多数のドメインに特化した語句を登録する必要があり，処理の効率化が求められる．

これに対し，本稿で提案する手法では複合語のうちドメイン関連度が一定値以下のものをフィルタリングする処理を行う．こうすることでドメインに関係のない複合語を除くことができるため，残った複合語は上位語が等しければ，意味的に大きく異なることがないと考える．その後，上位語が等しい語句を同一の語句として扱いながら属性表現の抽出を行うことで，従来の二手法の長所を組み合わせた，網羅性が高く，かつ精度の高い辞書の低コストでの作成を実現した．

3　提案手法

提案するシステムは，

＜1＞複合語の収集
＜2＞ドメイン関連度によるフィルタリング
＜3＞属性表現の抽出

の 3 つのモジュールから構成される．全体の構成を図 1 に示す．

3.1　複合語の収集

テキストデータを南瓜[3]で形態素解析し，複合語の収集を行う．複合語は形態素解析を行うと「名詞－固有名詞」＋「名詞－一般」のようにバラバラになってしまうため，再構成する必要がある．構成要素には南瓜の品詞体系で以下に該当する語句を用いる．

名詞－一般，名詞－サ変接続，名詞－固有名詞，記号－アルファベット

抽出した複合語には多くの造語，略語や形態素解析の誤り，書き手のミスによるノイズが含まれている．そこで収集された複合語 w に対して，それ自身をクエリとした Yahoo! Japan[4]の検索エンジンでのヒット数 h_w を求め，閾値 T_1 に関する(1)式を満たす語句を除去する．なお検索の際は，クエリをダブルクォーテーションでくくり，フレーズ検索を行う．

$$h_w < T_1 \quad (1)$$

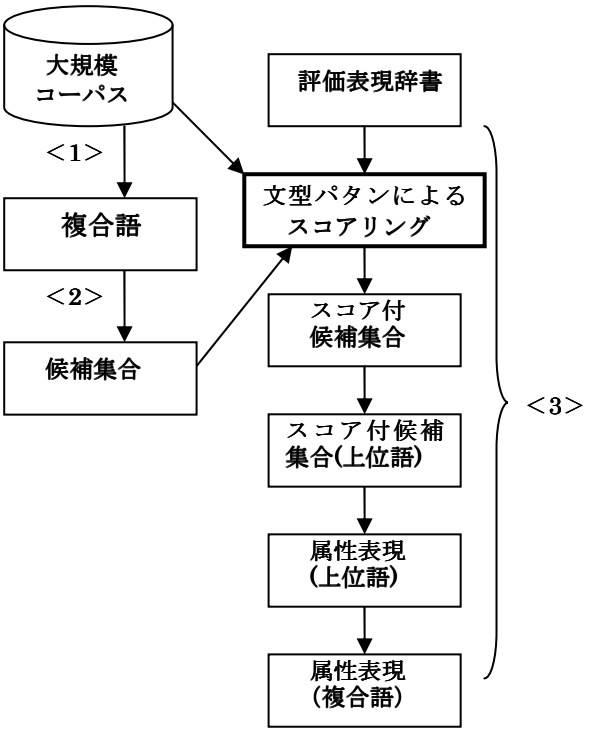

図 1，属性表現の抽出システム

3.2 ドメイン関連度によるフィルタリング

前節で収集した複合語うち，ドメインに対する関連度が低い語句，つまり属性表現となる可能性が低い語句のフィルタリングを行う．こうすることで，例えばプリンタドメインにおいて「高画質モード」,「フォトモード」などを残し,「i モード」,「留守電モード」などの属性表現として不適当な語句を除去することができる．残った語句は上位語が等しければ意味的には大きく異なることはないと考えられるので，これ以降まとめて処理することができる．ドメイン関連度 D は，語句 w をクエリとした検索ヒット数 h_w，ドメイン名 d をクエリとした検索ヒット数 h_d，及び語句 w とドメイン名 d の AND 検索をした場合の検索ヒット数 $h_{w,d}$ を用いて，(2)式で算出した．D を用いることで異なる w と d の組の関連度を同一尺度で扱えるため，(3)式で示す閾値 T_2 を異なるドメイン間で用いることができる．なお検索は，前節同様フレーズ検索を行い，$h_{w,d}$ に関しては w と d のフレーズ間の AND 検索を行う．

$$D = \log_2\left(\frac{h_{w,d}}{h_w \times h_d}\right) \tag{2}$$

次にドメイン依存度 D と閾値 T_2 に関する(3)式を満たす語句に対し，フィルタリングを行った．

$$D < T_2 \tag{3}$$

残った複合語の集合を候補集合と呼ぶ．

3.3 属性表現の抽出

候補集合から属性表現を抽出する．抽出は以下の 4 ステップを踏むことで実現される．以下，順に説明する．

ⅰ．文型パタンによるスコアリング

以下に示す小林ら[1]の提案した文型パタンと人手で作成した評価表現の辞書とを利用して，候補集合にスコアリングを行う．なお，評価表現の辞書は現代形容詞用法辞典[5]を参考にし，使用頻度が高く，かつ一般性がある語句を独自の判断で 250 表現用意した．

1. ＜属性表現＞が/は/も/に/を ［評価表現］

ex,〈カラーインク〉が[汚い]

2. ［評価表現］ ＜属性表現＞

ex,［美しい］＜印刷画質＞

例えば，「接続スピードが遅い」という書き込みがあったとする．今，「遅い」が評価表現辞書に登録されているとすると，上記 **1** に合致し，「接続スピード」のスコアに 1 が加算される．

ⅱ．候補集合の上位語への変換

複合語を上位語へ変換する処理を行う．複合語はそれを構成する語句の集合のうち一番最後の語句を上位語とする．つまり「フォトモード」「デジタル画質」はそれぞれ「モード」「画質」に変換される．その際，上位語 u のスコア S_u は u を上位語とする複合語の集合 C_u に含まれる複合語 w のスコア $score_w$ を用いて(4)式で表わされる．

$$S_u = \sum_{w \in C_u} score_w \tag{4}$$

例えば，「モード」を上位語とする複合語が「高画質モード」と「フォトモード」であり，前者がスコア 3 で，後者がスコア 6 であれば，「モード」のスコアは 9 になる．

ⅲ．属性表現（上位語）の抽出

上位語に変換された候補集合に対して，スコア S_w の閾値 T_3 に関する(5)式を満たす語句を属性表現として抽出する．

$$S_w > T_3 \tag{5}$$

抽出された語句を，手動で属性表現か否か分類する．抽出の際は，その語句を上位語とする複合語が属性表現としてふさわしいかどうかを判断基準とした．

ⅳ．属性表現の複合語への変換

ⅲで抽出された属性表現に関して，**ⅱ**で変換元になった表現を復元する．例えば「モード」の場合，「高画質モード」「フォトモード」等が復元される．これを属性辞書として登録する．

4 評価実験

大規模コーパスのテキストデータとして，価格.com 掲示板[6]からプリンタドメインの書き込み約 15 万文を取得した．できる限り精度よく，かつ網羅的に属性表現抽出を行えるように表 1 に示した数値に閾値を設定し，提案手法

を用いて実験を行った。実験の結果，各モジュールでの出力数は表 2 に示すようになった．なお 3.2 で示したドメイン名のクエリは「プリンタ」とした．

表 1，閾値

閾値名	T_1	T_2	T_3
閾値	300	−29.8	15

表 2，各モジュールの出力

	複合語	候補集合	属性表現（上位語）	属性表現（複合語）
数	19461	12138	234	1966

4.1 ドメインに無関係な複合語除去の精度・再現率

本手法で提案したドメイン関連度によるフィルタリングがどの程度有効性があったのか検証する．まずは，フィルタリングを行うことで，複合語の集合からどの程度精度よく網羅的に，プリンタドメインに関係のない複合語を除去出来ていたのかを調べた．ここで，プリンタドメインに関連のない複合語とは，「i モード」「ゲーム画面」のように，上位語が属性表現（上位語）として抽出されているにも関わらず，複合語になることで別のドメインに属する複合語のことを指す．上位語が属性表現（上位語）として抽出された複合語に関して，フィルタリングで残った複合語の集合 A の精度 p，再現率 r を(6)，(7)式を用いて求めた結果，それぞれ 93.1%，91.4%になった．なお，フィルタリングで残った複合語の集合 A のうち，属性表現（複合語）である語句の集合を B とし，登録すべき属性表現（複合語）の集合を C とする．

$$p = \frac{B}{A} \tag{6}$$

$$r = \frac{B}{C} \tag{7}$$

またフィルタリングを行わないで辞書を作成した場合，456 個の属性表現としてふさわしくない複合語を登録してしまうことがわかった．これら結果は，高い精度でフィルタリングに成功していることを示し，フィルタリングの有効性が示された．なお，後述の ii，iii，iv で述べるノイズはフィルタリングには無関係であるため除去して考えた．

次にフィルタリングの有効性が閾値 T_2 によってどう変化するのかを調べるため，精度，再現率と閾値 T_2 の関係を図 2 に示す．

図 2，フィルタリングで残った語句の精度と再現率

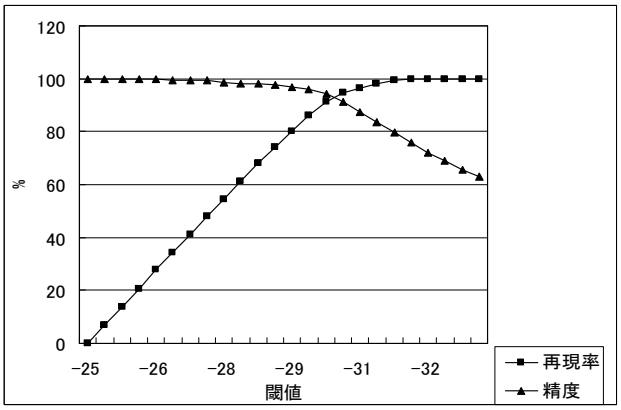

図 2 のグラフは，閾値を小さくすればするほど精度は低くなる一方，再現率は高くなる．また，閾値を大きくすればするほど，再現率は低くなるが精度は高くなる．図 2 のグラフより，T_2 を−29.8 と設定すると，再現率・精度両者ともに高い値を示す結果をなった．

4.2 フィルタリング誤りの原因

属性表現抽出対象のドメインに関連する複合語であるにも関わらず，提案手法によって誤ってフィルタリングされたケース(false positive)が発生した．また対象ドメインに関する複合語ではないのにも関わらず，ドメインに関連する複合語と判断されたケース(false negative)も発生した．ここでは，これらの誤りの原因を考察するため，具体例として属性表現（上位語）として抽出された「スピード」と「画面」を上位語とする複合語について，フィリタリング後に残った語句と除去された語句を表 3 にまとめた．なお，表 3 では「読み込みスピード」は「読み込み」と表記する．

表 3，フィルタリング語に残った語句と除去された語句

	残った語句	除去された語句
スピード	出力，コピー，スキャン，印刷，プリント，印字	読み込み，高速，通信，猛，高，開発，接続
画面	エラー，開始，トリミング，Web，初期，印刷，BIOS，モニタ，警告	管理，注文，送信，ゲーム，終了，接続，サンプル，選択，プリントアウト

フィリタリングを残った語句の中に「BIOS 画面」のようにプリンタに関連のない語句があった．また除去された語句の中にも「読み込みスピード」や「プリントアウト画面」のように属性表現として入れるべき語句がいくつかあった．前者の特徴として，語句をクエリとして検索された

Webページのうち,「プリンタ」と共起するページの割合は多いが，その語句と「プリンタ」が同一ページ内で異なるトピックで使われていることが多く,意味的な関連性はあまり深くないことがあげられる.また後者の特徴としては,「プリンタ」とWebページ内で共起する場合は，同一トピックに共起するなど関連性は非常に高いのだが,その語句自体の出現頻度も高いためドメイン関連度Dが低く算出されていることがあげられる.

以上のことから語句wとドメイン名を結合してクエリとした場合の検索ヒット数$h_{w,d}$は，ページ内にこれら二つのクエリが存在するページの数であるため,それらの語句がページの中でどのような関係性をもっているのかまでは考慮できていないことが本手法の問題点であると考えられる.

今後は検索ヒット数だけでなく,二つのクエリが出現するページ内で同一トピックの中で共起しているのかどうかなどを考慮した,より精密なドメイン依存度判定を行う必要があると考えられる

4.3 属性表現辞書の精度について

抽出した属性表現の辞書に関して精度を測ったところ,78.5%であった．ノイズは大まかに分けて以下のように分類された.

ⅰ．プリンタに関連性のない語句

「アンチウイルスソフト」や「BIOS画面」などのプリンタに関連性のない語句が抽出された.前節で述べたとおりドメイン関連度によるフィルタリングがうまくいかなかったためだと考えられる.

ⅱ．形態素解析の誤りによるノイズ

形態素解析の誤りによるノイズとして,「から印刷データ」や「あとインク」といった語句が抽出された．これらは「から」や「あと」などの本来助詞や接続詞と解析されるべき語句が名詞として認識されてしまうことが原因であった．語句としては不自然なものなのだが,「印刷データ」や「インク」自体の検索ヒット数が多いため，閾値を超える検索ヒット数をカウントしたものと考えられる.

こうした解析誤りが形態素解析器の性質上避けられない問題なのか確認するため,これらのノイズをクエリとしてWebページを収集し，その中からノイズが出現する文を抜き出し形態素解析を行った結果，ほとんどの文で「あと」や「から」などの接続詞や助詞に関して，正しく形態素解析が行われた．このことから，解析誤りは常に起こるものではなく,ある特殊な場合のみに起こることがわかる.よって各語句に関して，前述のようにWeb ページ上の文を利用して形態素解析をやり直すことで,解析誤りによるノイズを除去することが可能であると考えられる.

ⅲ．複合語の構成要素として選択しなかった品詞によって構成される語句

「再利用インク」の「再」（接頭詞-名詞接続）や「CD－R メディア」の「－」（記号－一般）などの構成要素として選択しなかった品詞によって構成される語句は「利用インク」や「Rメディア」のように意味を成さないノイズとして抽出されてしまう．しかし単純に「接頭詞-名詞接続」や「記号－一般」を構成要素にしてしまうとノイズも増えてしまうため,これらの品詞の中でも構成要素にして良い語句と,してはいけない語句を分類する必要があると考えた.

ⅳ．その他のノイズ

ドメインに関連性はあり,形態素解析の誤りもないのだが，辞書に登録するには不適当な語句があった．具体的には「各社」,「当社」,「上記」などの抽象的な語句よって修飾される語句である.

5 まとめ

本稿ではドメインごとの複合語の属性表現辞書を精度よく網羅的に低コストで抽出する手法を提案した.実験により精度よく低コストで複合語の属性表現を抽出できている事が示された.今後は属性表現辞書の網羅性について評価を行うことに加え，4.2で述べたようなノイズを減らして，抽出の精度をあげていく予定である.

参考文献

[1] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一(2003).“テキストマイニングによる評価表現の収集” 情報処理学会研究報告，NL154-12 pp77-84.
[2] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一(2005).“意見抽出のための評価表現の収集” 自然言語処理，12(2)，pp203-222.
[3] 日本語係り受け解析器“南瓜”
http://chasen.org/~taku/software/cabocha/
[4] Yahoo! JAPAN, http://www.yahoo.co.jp/
[5] 飛田良文，浅田秀子(1991).“現代形容詞用法辞典” 東京堂
[6] 価格.com 掲示板，http://kakaku.com/bbs/