評判情報分析のための製品属性の異表記辞書の自動構築

LIU Chaoyu　　白井 清昭
北陸先端科学技術大学院大学 情報科学研究科
{s1410047,kshirai}@jaist.ac.jp

1 はじめに

製品を対象とした評判情報分析においては，製品そのものに対するユーザの評価や意見だけでなく，製品の属性に対する評価の分析が求められることが多い．例えば，パソコンの場合，価格，サイズ，メモリ，ディスク容量，拡張性，キーボードなどが製品の属性に該当し，「価格は安いがメモリが少ない」「サイズは手頃だがキーボードは打ちにくい」といったように製品の属性に対する意見を調べたい．製品属性を対象とした評判情報分析でしばしば問題となるのは異表記 (表記ゆれ) である．異表記とは，同一の実体を指し示す複数の異なる表現を指す．例えば，「価格」は「値段」「売り値」「購入金額」「コスト」といった様々な表現で表わされる．ある製品の価格に関する評判を網羅的に収集するためには，これらの表現が全て「価格」の異表記であり，同じ製品属性を指すものであることを認識する必要がある．

本論文では，製品属性を対象とした評判情報分析のための基礎的な知識として，製品属性の異表記辞書を自動的に構築する手法について述べる [2]．ここでの製品属性の異表記辞書とは，「パソコン」「テレビ」「冷蔵庫」といった製品のカテゴリ毎に，製品の属性を異表記も含めて網羅的に収集した辞書と定義する．ウェブ上に存在する製品ページの仕様表ならびにユーザーレビューのテキストから，製品カテゴリに特化した異表記辞書を自動構築する．

2 関連研究

異表記の語をコーパスから自動獲得する研究として，カタカナで表記された外来語や翻字の異表記を検出する手法がいくつか提案されている．異表記の語の検出には 2 つの尺度が組み合わせて用いられる．ひとつは 2 つの単語間の編集距離，もう一つは大規模コーパスにおける単語の周辺文脈の類似度である [3, 5]．大前と黄瀬は，ウェブ上の表を解析し，属性と属性値を抽出する手法を提案している [4]．さらに，属性を，その属性が持つ属性値のベクトルとして表現し，類似したベクトルを持つ属性は同じ実体を表わすとみなして統合している．ただし，彼らは製品属性の抽出に目的を限定していない．Shinzato と Sekine は，オンラインショッピングサイトにおける商品説明文から製品の属性・属性値を自動抽出する手法を提案している [6]．まず，(1) 商品ページにおける表および箇条書きから属性と属性値の組を抽出し，(2) 商品説明文に対して属性と属性値を自動的にタグ付けし，(3) 属性・属性値抽出モデルを機械学習している．ステップ (1) において，同じ属性値を持つ属性は異表記であるという考え方に基づいて異表記の製品属性を同定している．

本研究は，製品の仕様表とテキストの両方から異表記の製品属性を抽出する点で Shinzato と Sekine の研究と類似している．ただし，本研究では，単一の通販サイトではなく複数のメーカーのウェブページの製品仕様表から属性を抽出する点，属性抽出の対象テキストとして商品説明文ではなくレビュー文を用いる点，レビュー文からは属性・属性値の組ではなく仕様表から抽出した属性の異表記を抽出する点に違いがある．

3 提案手法

ここでは，「パソコン」「冷蔵庫」のような製品カテゴリが入力として与えられたとき，そのカテゴリの製品の評価によく使われる属性を収録した製品属性異表記辞書 D を構築することを目的とする．本研究では D を以下のように表わす．

$$D = \{\cdots, A_i, \cdots\} \qquad (1)$$
$$A_i = \{\cdots, a_{ij}, \cdots\}$$

a_{ij} は製品の属性を表わす単語 (もしくは複合語) であり，集合 A_i は同じ実体を表す異表記の属性の集合である．A_i を集めたものを D とする．

提案手法の概要を図 1 に示す．提案手法は大きく 2 つに分けられる．ひとつは，製品のウェブページなどから製品の仕様表を抽出し，さらに仕様表から製品属性を抽出して，初期の製品属性異表記辞書 (D_t と記す) を構築する処理である．もう一つは，製品に対するレビュー文を知識源とし，製品属性を抽出するパターンを獲得し，そのパターンを用いたマッチングによって異表記の製品属性を獲得する処理である．

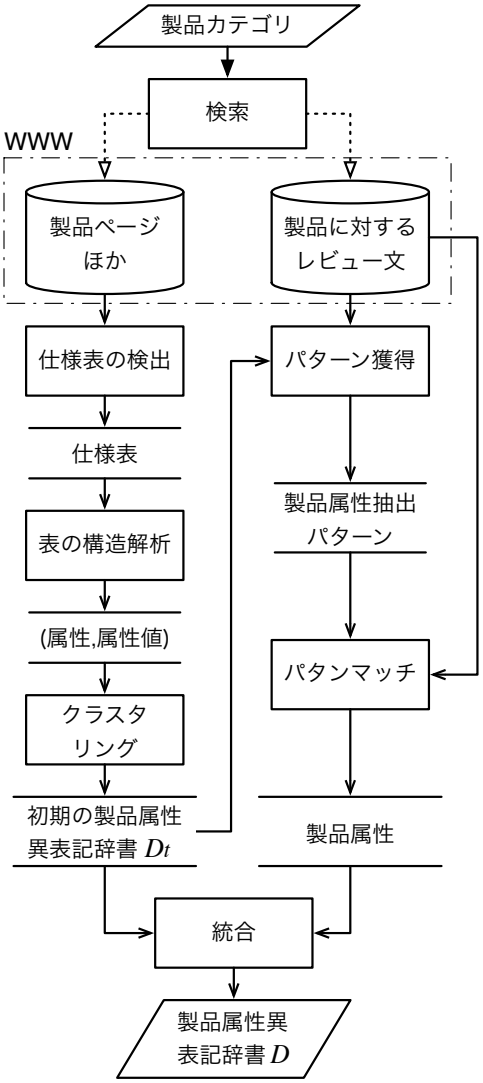

図 1: 提案手法の概要

3.1 仕様表からの属性抽出

ここでは製品の仕様表から製品属性を抽出する手法について述べる．

1. 価格.com [1] のサイトから，入力として与えられた製品カテゴリに該当する製品のページを取得する．
2. 価格.com の製品ページから仕様表を取得する [2]．
3. 価格.com の製品ページの中に含まれる「メーカー製品情報ページ」というリンクから，その製品のメーカーのウェブサイトへの URL を得る．
4. メーカーのウェブページから table タグでマークアップされた表のうち，図 2 の条件のいずれかを満たすものを取得する．

[1] http://kakaku.com/

[2] 価格.com では仕様表は ‘tblBorderGray’ で始まる class 属性を持つ table タグでマークアップされている．

- ⋆ table タグの class 属性が文字列 ‘spec’ を含む．
- ⋆ HTML の DOM において，table タグの直前に出現する兄弟ノードのタグの class 属性が文字列 ‘spec’ を含む．
- ⋆ そのページにおける唯一の table タグである．
- ⋆ table タグの直前に出現する兄弟ノードが「仕様」など仕様表を示唆するキーワードを含む．

図 2: 仕様表の条件

5. ステップ 2. と 4. で取得した表から属性と属性値の組を取得する．

価格.com から取得した仕様表はフォーマットが決まっているため，属性と属性値は容易に抽出できる．一方，メーカーのページに掲載されている仕様表からは，以下の手続きにしたがって属性と属性値の組を取得する．

- ○ 表の行に th タグが 0 個，td タグが 2 個以上ある
  → 最初の td から属性，残りの td から属性値を抽出
- ○ 表の行に th タグが 1 個，td タグが 1 個以上ある
  → th から属性，td から属性値を抽出
- ○ 表の行に th タグが 2 個，td タグが 1 個ある
  → 2 つの th から属性，td から属性値を抽出
- ○ 表の行に th タグが 2 個，td タグが 0 個ある
  → 最初の th から属性，次の th から属性値を抽出

次に，属性と属性値の組に対してクラスタリングを行う．ここでの目的は，異表記の属性をひとつにまとめることである．まず，(属性, 属性値) の組を素性ベクトルで表現する．素性ならびにその重みの一覧を表 1 に示す．A は属性，V は属性値から取得される素性を表わす．また,「属性値内の文字列」とは，属性値を区切り文字 (スペース, カンマ，括弧など) で区切った後，数字の列を ⟨N⟩ で置き換えた文字列である．

表 1: 属性，属性値の組の素性ベクトル

素性	重み
A 属性そのもの	10
A 文字の 3-gram, 2-gram, 1-gram	3,2,1
V 属性値内の文字列 (⟨N⟩，⟨N⟩+単位)	2
V 属性値内の文字列 (その他)	1

凝集型クラスタリングアルゴリズムによってクラスタリングを行う．クラスタ数は (属性, 属性値) の組の総数の 90% と設定する．クラスタリング後，属性値を取り除いて，初期の製品属性異表記辞書 D_t を得る．

表 2: 実験データ

製品カテゴリ	PC	カメラ	テレビ	腕時計	冷蔵庫	炊飯器	洗濯機	レンジ	エアコ
メーカー数	6	4	8	2	3	6	4	3	1
レビュー文数	240,690	327,544	280,790	109,980	44,042	45,547	73,951	31,424	37,450

PC: ノートパソコン　カメラ: デジタル一眼カメラ　レンジ: 電子レンジ　エアコ: エアコン

このとき，式 (1) において，クラスタが A_i に，クラスタ内に属する属性が a_{ij} に該当する．

3.2 レビュー文からの属性抽出

製品カテゴリに関する製品のレビュー文から異表記の製品属性を獲得する手法について述べる．まず，製品属性を抽出するパターンの候補を獲得する．パターンのテンプレートは以下の通りである．

$$\text{[A]}\ w_1 \cdots w_l\ \text{[E]} \quad \text{または} \quad \text{[E]}\ w_1 \cdots w_l\ \text{[A]} \qquad (2)$$

[A] は初期の製品属性異表記辞書に含まれる属性，[E] は評価語である．評価語は日本語評価極性辞書 (用言編)[1] に登録されている語とする．w_i は [A] と [E] の間に出現する単語の列で，l は 1,2,3 のいずれかとする．レビュー文から [A] と [E] が近くに出現する文を検索し，上記のテンプレートを用いてパターンの候補を生成する．

獲得したパターンの候補 P_i のスコアを式 (3) のように定義する．すなわち，初期の製品属性異表記辞書 D_t の中に含まれる属性を多く抽出できるパターンほど信頼性が高いとみなす．

$$S(P_i) = \frac{\text{パターン } P_i \text{ にマッチしかつ } D_t \text{中の属性が抽出される文の数}}{\text{パターン } P_i \text{ にマッチする文の数}} \qquad (3)$$

マッチする文の数が 3 以上で，かつ $S(P_i)$ が 0.5 以上のパターンを製品属性抽出パターンとして獲得する．得られたパターンをレビュー文集合に適用し，新たな製品属性を得る [3]．

3.3 製品属性異表記辞書の構築

3.1 項で獲得した初期の製品属性異表記辞書 D_t と，3.2 項の方法で獲得した製品属性を統合して，最終的な製品属性異表記辞書を得る．統合は，パターンマッチで獲得した属性を D_t に併合することで実現する．パターン P_j によって得られる属性のうち，D_t に既に含まれている属性の集合を K_j，含まれていない属性の集合を U_j とおく．K_j の要素は D_t における属性集合 A_i のいずれかに属する．K_j の中で出現頻度が最大の A_i を求め，P_j は A_i の属性の異表記を抽出するためのパターンとみなす．そして，U_j を A_i に追加する．

[3] 計算時間を短縮するため，スコアの計算はレビュー文集合の一部のみを用いる．そのため，$S(P_i) = 1$ の場合でも，D_t に含まれない新たな製品属性が得られる可能性がある．

4 評価実験

評価対象として設定した 9 つの製品カテゴリを表 2 に示す．表 2 には，各カテゴリ毎に，仕様表を抽出するために用いたメーカーのウェブサイトの数，ならびにレビュー文の数も示している．レビュー文は価格.com のサイトから収集した．

4.1 仕様表からの属性抽出の評価

表 3 に，抽出された仕様表の数，ならびに仕様表から抽出された (属性, 属性値) の組の数を示す．これは 9 つの製品カテゴリの合計である．メーカーページについては，仕様表抽出の精度と再現率，ならびに (属性, 属性値) の抽出の精度も示した．(属性, 属性値) の抽出精度は，ランダムに選択した 100 組に対して算出した．表 3 より，仕様表も (属性, 属性値) の組も抽出精度は十分に高いことがわかる．

(属性, 属性値) の組のクラスタリング結果を評価する．ここでは式 (4) に示す Purity を評価基準とする．majority(A_i) は A_i の中で同一実体を指す属性の最大値である．また，C は，作成されたクラスタのうち，$|A_i| > 1$ であるクラスタの集合である．すなわち，要素数 1 のクラスタは評価から除外した．

$$\text{Purity} = \sum_{A_i \in C} \frac{|A_i|}{|C|} \cdot \frac{\text{majority}(A_i)}{|A_i|} \qquad (4)$$

結果を表 4 に示す．「全カテゴリ」は 9 つの製品カテゴリについての合計，「平均」はそれらの平均である．ク

表 3: 仕様表の抽出・仕様表からの属性の抽出の評価

		仕様表	(属性, 属性値)
価格.com	抽出数	37	972
メーカーページ	抽出数	103	1027
	精度	0.89	0.90
	再現率	0.82	—

表 4: 製品属性のクラスタリングの評価

	全カテゴリ	平均
クラスタ数	880	97.8
$\|A_i\| > 1$ のクラスタ数	101	11.2
Purity	0.829	0.790

ラスタリングの Purity は，全体で 0.829 と高く，製品カテゴリ毎に見ても,「腕時計」で 0.500,「エアコン」で 0.667 であるが，それ以外のカテゴリでは 0.75 以上と高い値が得られている．一方，要素数が 1 より大きいクラスタの数が少ないことから，異表記の属性が同じクラスタにまとめられていない可能性がある．クラスタリングの際に設定するクラスタの数を小さくすればより大きなクラスタを構築できるが，Purity は低下するだろう．現在はクラスタ数は全属性数の 90%と設定しているが，今後は最適化なクラスタ数を決める方法を探究する必要がある．

4.2 レビュー文からの属性抽出の評価

レビュー文からのパターン獲得ならびにパターンマッチによって獲得された異表記の属性を評価する．ここでは,「PC」「カメラ」「テレビ」の 3 つのカテゴリのみを評価対象とする．表 5 は，獲得されたパターンの数，パターンによって抽出された属性の数，そのうち初期の製品属性異表記辞書に登録されていない (新たに獲得できた) 属性の数，そのうち正しい異表記の属性とみなせるものの数，及び抽出精度を示している．抽出精度の定義は式 (5) の通りである．

$$\text{抽出精度} = \frac{\text{異表記とみなせる属性数}}{D_t\text{に含まれない属性数}} \qquad (5)$$

抽出精度は低く，改善の余地がある．現時点では，パターンのテンプレートは，式 (2) に示すように，属性と評価語およびその間の単語という単純なものしか採用していない．属性抽出のための条件をより精密に定義できるようなテンプレートを用意することで，抽出精度を改善できると考えている．

表 5 に示した 3 つ以外の製品カテゴリについては，レビュー文の数が十分に多くないため，パターンマッチによる属性抽出の評価は行わなかった．今後，より多くのレビュー文を収集し，提案手法の評価を行う予定である．

4.3 獲得された属性の例

獲得されたパターン，およびそのパターンマッチによって獲得された異表記の属性の例を図 3 に示す．*が付いている属性は D_t にも含まれている属性を表わす．

表 5: パターンマッチによる属性抽出の評価

	PC	カメラ	テレビ
パターン数	27	16	5
抽出属性数	108	64	65
D_t に含まれない属性数	69	45	58
異表記とみなせる属性数	26	3	7
抽出精度	0.34	0.067	0.12

最初の例では,「キーボード」の異表記として,「キー」や「タイピング」が得られている．これらを含む意見文はキーボードに対する評価を表わすとみなせる．

パターン:	[A] の 感触 も [E]
製品属性:	キーボード*、キー、タイピング
パターン:	[A] の 持ち も [E]
製品属性:	バッテリー*、バッテリ、電池

図 3: 獲得されたパターンと製品属性の例

5 おわりに

本論文では，製品の仕様表ならびにレビュー文から異表記の製品属性を抽出し，製品属性異表記辞書を自動構築する手法について述べた．仕様表からの属性抽出の精度は十分高く，初期の辞書を構築する手法として有望である．一方，レビュー文から異表記の属性を抽出する手法は，獲得した属性数，抽出精度のいずれの観点からも十分ではない．4 節で議論した問題点を克服することが今後の課題となる．

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. 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, 2005.

[2] Chaoyu LIU. 評判情報分析のための製品属性異表記辞書の自動構築. Master's thesis, 北陸先端科学技術大学院大学, 3 2016.

[3] Takeshi Masuyama, Satoshi Sekine, and Hiroshi Nakagawa. Automatic construction of japanese KATAKANA variant list from large corpus. In *Proceedings of COLING*, pp. 1214–1219, 2004.

[4] 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, 2006-NL-171, Vol. 2006, No. 1, pp. 43–48, 2006.

[5] Kiyonori Ohtake, Youichi Sekiguchi, and Kazuhide Yamamoto. Detecting transliterated orthographic variants via two similarity metrics. In *Proceedings of COLING*, pp. 709–715, 2004.

[6] Keiji Shinzato and Satoshi Sekine. Unsupervised extraction of attributes and their values from product description. In *Proceedings of IJCNLP*, pp. 1339–1347, 2013.