商品説明文に対する属性値タギング

宇佐美 佑　　萩原 正人　　関根 聡
東京大学大学院情報理工学系研究科　　楽天技術研究所
yusmi@is.s.u-tokyo.ac.jp
{masato.hagiwara, satoshi.b.sekine}@mail.rakuten.com

1　序論

ショッピングサイトの普及と大規模化が進む現在，登録された商品の “検索しやすさ” が売り手と買い手（ユーザ）双方にとって重要である．ユーザの多岐に渡るニーズを満たすため，商品の絞り込み検索，いわゆるファセットサーチの実現は必要不可欠である．ここで絞り込みの手がかりとなるのは，たとえば服飾商品におけるサイズがMであるとか，色が青色であるといった，他の服飾商品と共通する性質（サイズ，色）の，具体的な値（M，青）である．これらを個々の商品が持つ属性と捉えることにより，属性が各商品を特徴付けていると考えられる．本稿では性質を現す表現（サイズ，色など）を属性名，具体的な値（M，青など）を属性値と定義し，これらのペアを商品の持つ属性とする．

多くのショッピングサイトでは属性を手作業で入力しているが，これはコストが大きい．そこで商品説明文を用い，自動的に商品と属性を結びつける手法に取り組んだ．具体的には，商品への属性の紐付けをその商品説明文中の属性値表現への属性名ラベル付与問題（属性値タギング）として扱い，固有表現抽出と同様に対象表現の辞書を利用し [3　] 以下の手順で解く．

商品説明文より，対象属性名に対応する属性値辞書を構築する．

2　少量の商品説明文に対し，人手で属性値をアノテートし，コーパスを作成する．

3　構築した属性値辞書と作成したコーパスを用いて属性値タガーの訓練をし，それによって商品説明文に対して属性値タギングする．

実験と評価には，楽天市場ワインカテゴリに含まれる全商品（2　　5　点）を用いた[1]．

本研究で得られた知見は以下の通りである．

構築した辞書を利用することで，固有表現抽出と同様に属性値タガーの性能向上が認められた．

網羅性の高い辞書を構築しても，局所的な曖昧性を含む表現などにより，属性値タガーが高性能にならない可能性がある．

2　関連研究

本研究に最も近いのは，商品の属性名，属性値を抽出する試みである．G ani ら [ ] や Prob t ら [ ] は，まず属性名，属性値のペアをシードとして数種類用意した．それらの商品説明文中での出現パターンを利用し，ブートストラップ法で属性名，属性値の新たな表現を抽出した．Putt ivid ya ら [ ] は商品説明文よりさらに短い，商品タイトルのみを用いている．これらは全て属性名と属性値の辞書拡張のみを目的としており，本研究の目指す各商品への属性の紐付けはこれらの次の段階にあたる．

属性抽出が直接の目的ではないものの，　iu ら [5] や Pope cu ら [ ] は商品についての評判分析をする際に，まず商品の特徴として属性名を抽出しその属性名への言及を評判分析の手がかりとした．　rody ら [2] は，ユーザの入力する商品レビューを様々な側面で解析するために，　D　を用いて局所的な話題を識別した．抽出された話題と特徴語の関係には，属性名，属性値の関係も含まれた．

3　提案手法

3.1　辞書構築

まず，対象属性名に対応する属性値辞書を構築する．対象となる属性名は，商品説明文より自動抽出 [　] し

[1] 楽天データ公開にて公開されているデータを利用した．http://rit.rakuten.co.jp/rdr/

表　ワインカテゴリにて使用した属性名ラベル，属性名，対応する属性値例と作成したコーパス中の事例数

属性名ラベル	属性名	属性値例	事例数
alcohol	アルコール度数，アルコール分，アルコール度	14%，12.5 度	31
aroma	香り	スイカ，熟れたイチゴ	103
color	色	ルビー色，濃い紫色	17
cultivation area	栽培面積	40ha，400 ヘクタール	4
dish	合う料理	ステーキ，イタリア料理	39
grape	品種，ブドウ品種，葡萄品種，ぶどう品種，使用品種	CHARDONNAY，甲州種	261
parker point	パーカーポイント	88 点，100 点	13
production amount	生産量，年間生産量，生産本数	2953 ケース	1
production area	産地，原産国，生産地	シャンパーニュ地方，ドイツ	235
rating	格付，格付け	グランクリュ，AOC	37
temperature	飲み頃温度	16-20 度，10 ℃	4
soil	土壌	火山土壌，貝殻石灰岩	7
taste	味わい，味のタイプ	辛口，ミディアムボディ	96
tree age	平均樹齢，樹齢	42 年，30 年以上	4
type	種類	ロゼ，赤ワイン，白泡	158
volume	容量，内容量	750ML，375ML	90
winery	生産者，ワイナリー名	DOMAINE ASTRUC，グラハム・ベック	183
year	ヴィンテージ，収穫年，生産年，醸造年	1991，NV，2007 年	130

た属性名候補から人手で選別し決定した．それぞれの属性名には，後の属性値タギング時にクラスとして用いる属性名ラベルを付与した．ワインカテゴリにて使用した属性名ラベルと属性名を表　に示す．

商品説明文において，属性値は対応する属性名の直後に隣接して出現する [　] ことを利用し，属性値を抽出する．ワインカテゴリ中の全商品説明文に対し，次の前処理をする（図　参照）．

数字表現は個々の値が重要ではないので，桁数を維持したまま 0 に置換する．次のステップで数字+単位が　トークンとなるように，数字と 2 文字以上のアルファベットが連続する直後にスペースを追加する．

2　記号[2]で文を区切る．区切られた各部分を　トークンとする．

3　トークンが対象属性名に一致する場合，属性名ラベルに置換する．

前処理ののち，ワインカテゴリ中の全商品説明文における各トークンの出現頻度と，各トークンが属性名ラベルの直後に出現する頻度をカウントする．これらより，各トークンの出現確率と属性名ラベルとの同時出現確率を推定する．得られた確率を用い，帰無仮説 “それぞれのトークンは独立して出現する” として　5　信頼区間で t 検定を行う[3]．帰無仮説が棄却できる場合，属性名ラベルの直後に出現するトークンは，その属性名に対応する属性値とみなし，属性値候補に加える．

[2]ひらがな，カタカナ，漢字，アルファベット，カタカナ間の中黒,「,」,「%」,「‘」,「&」,「〜」以外の文字

[3]各トークンの出現頻度は正規分布に従う，という仮定も含む．この点については今後確認する必要がある．

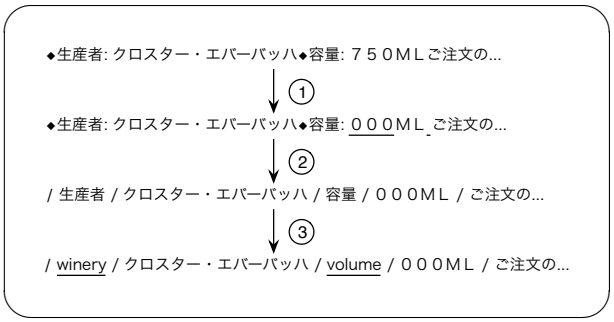

図　属性値辞書構築のための前処理実行例

このようにして得られた属性値候補 3 52　件を全て人手で確認し，属性名と対応のとれているトークンを属性値辞書に加えた．この結果 2　　件（　　）の表現が辞書のエントリとして得られた．この確認作業は，2 人のアノテータで　日間程度要した．

3.2　コーパス作成

コーパス作成のため，ワインカテゴリ中の全商品より　2 商品を無作為に抽出した．それらの商品説明文（タイトル含む）中の属性値表現に対し人手で属性名ラベルを付与し，タグ付きコーパスを作成した．対象属性名は表　と同じ　種類である．作成したコーパスに含まれる事例数を表　の右端に示す．コーパスの作成は 2 人のアノテータにより 2 日間で完了した．

3.3　属性値タガーの学習

前節において作成したコーパスを用いて　2 正則化 CRF モデルを訓練し，属性値タガーを開発した．CRF

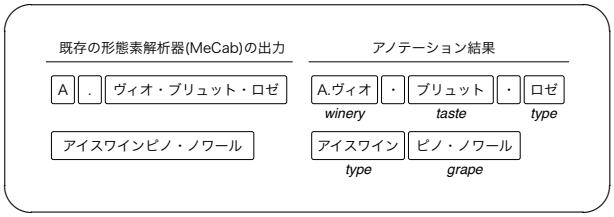

図 2 既存の形態素解析器による単語分割とアノテーション結果の違い

位置	c	pos	type	dic
$i-5$	厳	B-名詞-サ変接	C	O
$i-4$	選	I-名詞-サ変接	C	O
$i-3$	6	B-名詞-数	N	O
$i-2$	本	B-名詞-接尾	C	O
$i-1$	赤	B-名詞-一般	C	B-TYPE
i	ワ	I-名詞-一般	K	I-TYPE
$i+1$	イ	I-名詞-一般	K	I-TYPE
$i+2$	ン	I-名詞-一般	K	I-TYPE
$i+3$	セ	B-名詞-一般	K	O
$i+4$	ッ	I-名詞-一般	K	O
$i+5$	ト	I-名詞-一般	K	O

図 3 位置 i にて利用される素性の例

の実装としては CRFSuite[4]を用いた．商品説明文には未知のカタカナ語が非常に多く含まれている．そのため，図 2 のように既存の形態素解析器では正しく単語分割ができず，アノテーション結果と単語区切りが一致しない．この問題を避けるため，単語ベースではなく文字ベース [ ] を採用した．

学習とタギングの際に利用する素性は，前後 5 文字分の文字（c），品詞（po ），文字種（type），辞書素性（dic）である．品詞は MeCab[5]を用いて単語ごとに付与した後，文字ごとに　　2 記法にて分割した．また，辞書素性は構築した属性値辞書との最長一致文字列に　　2 記法にて付与した．位置 i にて利用される素性の例を図 3 に示す．

4 実験と評価

構築した辞書の効果と，辞書の網羅性が性能に与える効果を評価するため，次の 3 種の設定で実験した．

構築した辞書を用いない教師あり学習（S）

構築した辞書を用いた教師あり学習（S+D）

構築した辞書に，コーパスに含まれる属性値表現を全て追加したものを用いた教師あり学習（S+D+　）

[4]http://www.chokkan.org/software/crfsuite/
[5]http://mecab.sourceforge.net/

表 2 各手法の評価結果

手法	P	R	F_1
S	55	5	3
S+D		5	
S+D+		5	2

実験には 3 2 節で作成したコーパスを用いた．ただし，アノテートされた事例数が 5 未満の属性については訓練に充分な量でないと判断し，タグを取り除いた．また，S，S+D，S+D+　については leave one out 交差検定法にて評価した．

各手法における適合率（P），再現率（R），それらの調和平均（F_1）を表 2 に示す．構築した辞書を素性に用いる（S+D）ことで，適合率が　　と，辞書を用いない場合（S）に比べ若干減少するが，再現率が 5　　，F_1 値が　　に改善した．また，すべての属性値表現が網羅的に辞書に含まれている場合（S+D+　）の適合率は　　，再現率は　5　，F_1 値は　2　であった．以上より，辞書素性が有効であることがわかる．

S+D+　の評価結果は，網羅的な辞書が構築できたとした場合にも，属性値タガーは高性能にはならないことを示唆している．S+D+　において，属性値タガーの予測が間違った例を図　に示す．例　では，ワインに合う料理を説明する文中に現れる “タイ” が，production area と誤って予測されている．実際にタイで生産されているワインも多く，production area 辞書には “タイ” が含まれており，これが原因で予測されてしまっていると考える．このように，実際の商品への言及とは異なる文脈において，辞書に一致する表現が fal e po itive として予測される例が多く，適合率を下げる主要因となっている．例 2 では，“シャトー・ブラン・ド・ブラン” が誤って winery と予測されている．winery の辞書に “シャトー” で始まる表現が多く含まれるため，“シャトー・ブラン・ド・ブラン” が商品名にも関わらず winery と判断されてしまったと考えられる．例 3 では，“ヴィーニョ・ヴェルデ” が type と予測されている．この表現は production area，type どちらの辞書にも含まれているため辞書素性が競合し，さらに “ヴェルデ” が type として訓練データに含まれるため production area と予測し難くなってしまっている．例　は “ピュリニー・モンラッシェ 1 級” が rating 辞書に含まれているのにも関わらず，production area と誤って予測している．これは，“1 級” という rating に特徴的な語が “ピュリニー” から

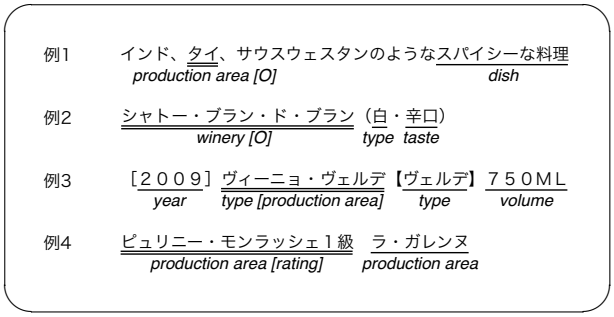

図　属性値タガーの間違い例（角括弧中が正解）

離れており，文脈として考慮されなかったことが原因と考えられる．また，コーパス中に “ピュリニー・モンラッシェ” が production area として多く出現したことも一因と考えられる．

これらの誤りを減らす手段として，局所的な素性のみでなくより広域の文脈情報を素性に組み込む方法が考えられる．そのひとつに，文章全体を考慮した大域的素性 [　] の導入がある．商品説明文全体にフランスの地名が production area として何度も出現する商品では，それを素性とすることで，例　のようにタイが production area と予測される可能性を下げることが考えられる．大域的素性でなくとも，文脈をさらに広く考慮したり，遠く離れた表現の情報を上手く素性に利用することが重要である．例 2 のように一部が特定の属性値表現に頻出する場合にも，文脈を広く考慮すれば頻出でない文字列部分まで捉えられ，誤った予測を減らすことが期待できる．このような手法を導入することで，例 3,　のように局所的な素性のみでは曖昧性を解決できない問題にも，対処できるのではないかと考える．

5　結論

本研究では，自動的に商品と属性を紐付けるための，商品説明文に対する属性値タギング手法を提案した．固有表現抽出と同様に辞書構築を行い素性として利用することで，単純な教師あり学習より性能が改善することを示した．その一方で，たとえ網羅的な辞書を構築しても，表現の曖昧性問題が解消せず，性能が十分に高くならないこともわかった．

今後は前節で議論した方針をもとに，性能改善，人的コストの削減，ワインカテゴリ以外の商品カテゴリへの拡張が課題になると考える．本手法は，ワインカテゴリに依存せず，全商品カテゴリに拡張可能である．しかし，現状は人的コストとして，属性名リストの選別，属性値辞書の選別，コーパスの作成が含まれており，全カテゴリで適用する場合の人的コストは高くなってしまう．現状のワインカテゴリのコーパスを用いた半教師あり学習によって他カテゴリへの分野適応をするなど，コストを抑える方法を模索したい．

今回の実験においては，商品説明文中の表現毎に予測の正解，不正解を判定していた．商品にどの属性が紐付くのかという問題には，属性値タギングをした後に多数決で代表属性を決定する方法なども考えられる．今後はこのような枠組みでの評価方法も検討していきたい．

参考文献

[1] Masayuki Asahara and Yuji Matsumoto. Japanese Named Entity extraction with redundant morphological analysis. In *Proceedings of HLT - NAACL '03*, pp. 8–15, May 2003.

[2] Samuel Brody and Noemie Elhadad. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *Proceedings of HLT - NAACL '10*, pp. 804–812. Association for Computational Linguistics, 2010.

[3] William W. Cohen and Sunita Sarawagi. Exploiting dictionaries in named entity extraction. In *Proceedings of ACM SIGKDD '04*, pp. 89–98, 2004.

[4] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, and Andrew Fano. Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, Vol. 8, No. 1, pp. 41–48, 2006.

[5] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of WWW '05*, pp. 342–351, 2005.

[6] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of HLT - EMNLP '05*, pp. 339–346, 2005.

[7] Katharina Probst, Rayid Ghani, Marko Krema, Andrew Fano, and Y. Liu. Semi-supervised learning of attribute-value pairs from product descriptions. In *Proceedings of IJCAI '07*, pp. 2838–2843, 2007.

[8] D.P. Putthividhya and J. Hu. Bootstrapped Named Entity Recognition for Product Attribute Extraction. In *Proceedings of EMNLP '11*, pp. 1557–1567, 2011.

[9] Yoshimasa Tsuruoka and Jun'ichi Tsujii. Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In *Proceedings of ACL '03 Workshop on NLP in Biomedicine*, pp. 41–48, 2003.

[10] 風間淳一, 鳥澤健太郎. 大域的素性を用いたタグ付けのためのパーセプトロン学習. 言語処理学会第 13 回年次大会発表論文集, pp. 103–106, 2007.

[11] 坂地泰紀, 小林暁雄, 関根聡, 竹中孝真. 商品ページからの属性・属性値抽出と同一商品クラスタリング手法. 言語処理学会第 16 回年次大会発表論文集, pp. 371–374, 2010.