応用

3.10 情報抽出

―商品の属性値抽出タスクのエラー分析―

新里 圭司（楽天技術研究所）

商品の属性値抽出タスク

場所・時間を気にせずに買い物可能なオンラインショッピングサイトは重要なライフラインになりつつある．商品の説明はテキスト形式で出店者より提供されるため，商品の属性－属性値を説明文から抽出し構造化された商品データを作成する属性値抽出技術のニーズは，オンラインショッピングサイトの現場ではきわめて高い．ここで「商品説明文から商品の属性値を抽出する」とは，たとえばワインの説明文「フランス産のシャルドネを配した辛口ワイン」から，（生産地，フランス），（ぶどう品種，シャルドネ），（タイプ，辛口）といった属性と属性値の組を抽出することを指す．商品の属性－属性値は，商品検索やレコメンド，マーケティング分析などさまざまな場面で活用できる．

ここでは商品の属性値抽出タスクを対象に行ったエラー分析について報告する．

分析対象データ

楽天データ公開☆1を通して配布されている商品データから，ワイン，シャンプー，プリンタインク，Tシャツ，キャットフードカテゴリに登録されている商品ページを無作為に20件ずつ，計100件抽出した．そして抽出したページをブロック要素タグ，「。」「？」「！」などの記号を手がかりに文に分割した．続いて，各商品ページのタイトル，商品説明文，販売方法別説明文に含まれる属性値を1名の作業者によりアノテーションした．カテゴリごとの対象属性およびアノテーションの規模を**表-1**に示す．このデータに対して，次章で述べる抽出システムを走らせ，システムの出力と人手アノテーションの差分を見ることでエラー分析を行う．

☆1 http://rit.rakuten.co.jp/opendataj.html

属性値抽出システム

タスクに内在する研究課題を明らかにするためのいち手段として，どのように動くかがグラスボックス的に分析できるシンプルなシステムを実行し，その結果を基に課題を明らかにする方法がある．今回はこのエラー分析方法にならい，事前に作成した属性－属性値辞書に基づいて属性値を抽出するシンプルなシステム（**図-1**）を作成し，このシステムのエラーの分析を行った．これは，今回の商品属性値抽出タスクは標準的なタグ付きコーパスや属性値抽出のためのソフトウェア等が公開されているわけではないため，複数のシステムを同一のデータセットで実行し，その差分を調査するといった分析が困難なためである．

属性－属性値辞書は商品ページに含まれる表や箇条書きなどの半構造化データから自動構築した．構築した辞書を人手で評価したところ，カテゴリによって差があるものの，その精度は平均で72.3%であった．属性値抽出の際は，まず入力文を形態素解析し，属性－属性値辞書中の属性値と最長一致した形態素列を対応する属性の値として抽出する．この辞書マッチに基づくシステムは，辞書に属性値が登録されているか否かで属性値の抽出を行うためエラーの原因の特定が容易である．

エラー分析

属性値の抽出は商品ページ内の半構造化データより自動構築した辞書に基づいている．辞書には誤った属性－属性値の組も含まれており☆2，これらに起

☆2 たとえばTシャツカテゴリに対する（色，8色）など．

カテゴリ	対象属性	文数	属性値数
ワイン	品種，容量，産地，生産者，タイプ，度数	365	262
シャンプー	容量, メーカ, 製造国, 成分, 商品名，サイズ，重量	638	490
インク	容量, サイズ, カラー, 重量, 適応機種，製造国	286	375
T シャツ	ブランド, サイズ, 素材, 色, 着丈，身幅，肩幅	720	357
キャットフード	メーカ，内容量，原産国, 粗繊維，粗脂肪，粗灰分, 水分，粗タンパク質	382	132
	合計	2,381	1,702

表 -1　対象カテゴリ，対象属性および対象データの規模．商品ページ数は各カテゴリともに 20 件

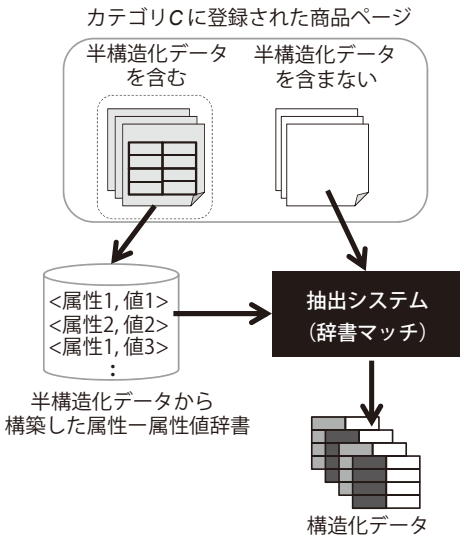

図 -1　属性値抽出の流れ．「半構造化データ」とは，表や箇条書きで記述された商品説明を指す

a. その他のシャンパン タイプ & スパーク & ワイン関連はコチラをクリック♪

b. 成猫体重 1kg 内容量 当り 1 日約 1.4 袋を目安として、1 日の給与量を 2 回以上に分けて与えてください。

c. アメリカ 製造国 各種機関で厳しい環境基準をクリアした分解作用で汚れだけを分解してくれるから髪や頭皮を傷めません。

図 -2　属性値を抽出する " 場所 " を誤った例

因する誤りが最も多く，誤抽出全体の 67.4% であった．自明だが，高い精度で辞書を構築することが辞書ベースの情報抽出システムにおいて重要であることが分かる．これはドメイン固有の知識を高精度で獲得する技術が求められることを意味する．

次に多かった誤りは属性値を抽出する“場所”に関連したものであった．商品ページ内には別商品へのナビゲーション，ブランドの説明や商品の使い方などさまざまなテキストが含まれており，当該商品の説明以外のテキストから抽出すれば，正しい属性ー属性値に基づいたものであっても誤る可能性が高い．**図 -2** に誤抽出の例を示す．図中の各事例は別商品へのナビゲーション（a），当該商品の関連情報（b），商品説明文中の名詞句の一部分（c）から抽出されたものである．このような誤りを減らすためには，商品ページのどの部分から属性値を抽出するべきかを認識する技術が求められる．

今後の展開

誤った辞書エントリに基づく誤抽出が最も多かったことから辞書の品質の向上が重要である．この課題に対しては，語彙統語パターンなどほかの手がかりとの組合せ，クラウドソーシングによる人手チェックの導入などで改善が期待できる．

2 つ目の課題「抽出する場所の誤り」については，「同じ出店者の複数の商品ページに含まれるテキストは当該商品との関連性が弱い」という仮定のもと，商品ページのテンプレートを認識することで，属性値を抽出するべき場所を特定する手法を考えている．

（2015 年 9 月 30 日受付）

新里 圭司（正会員） keiji.shinzato@rakuten.com
2006 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（情報科学）．京都大学大学院情報学研究科特任助教を経て，2011 年から楽天技術研究所．自然言語処理，特に情報抽出，評判分析の研究に従事．