情報抽出タスクの誤り分析 –商品の属性値抽出を題材に–

新里 圭司
楽天技術研究所
keiji.shinzato@mail.rakuten.com

1 はじめに

本稿では Project Next NLP 情報抽出タスクについて報告する．本活動の目的は情報抽出システムが抽出した結果の false-positve / negative の分析を通して，システムの性能向上にはどんな処理・データが必要かを明らかにすることである．情報抽出タスクとしては，実用面を考慮し，オンラインショッピングサイト上の商品説明文からの商品属性値の抽出を設定した．例えばワインカテゴリであれば以下の文が入力された時，(生産地，フランス)，(ぶどう品種，シャルドネ)，(タイプ，辛口) を抽出することを目的としている．

- フランス産のシャルドネを配した辛口ワイン．

商品の属性値抽出は実世界でのニーズが高く，実現できれば詳細なマーケティング分析（例えば「30 代女性にフランス産の辛口ワインが売れている」等），商品のレコメンド，ファセット検索などに利用できる．

2 分析対象データ

楽天データ公開 [1] より配布されている商品データから，論文 [3] を参考に，ワイン，シャンプー，プリンターインク，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，記号 [2] を手がかりに文に分割した．

カテゴリ毎に分析対象とした属性を表 1 に示す．これらの属性は論文 [3] で抽出対象とされたものに以下の修正を加えたものである．

- 同じ意味を表す属性名を人手で統合した．
- 誤った属性を人手で削除した．

[1] http://rit.rakuten.co.jp/opendataj.html

[2] 【, 】, 。, ？, ！, ♪, ※, ●, ○, ◎, ★, ☆, ■, □, ▼, ▽, ▲, △, ◆, ◇, ≪, ≫, ≪. ただし，これらが括弧内 (「」,『』) に出現している場合は区切らない．

表 1: カテゴリと分析対象属性

カテゴリ (ID)	対象属性
ワイン (100317)	品種，容量，産地，生産者，タイプ，度数
シャンプー (210677)	容量，メーカー，製造国，成分，商品名，サイズ，重量
プリンターインク (502185)	容量，サイズ，カラー，重量，適応機種，製造国
T シャツ (551180)	ブランド，サイズ，素材，色，着丈，身幅，肩幅
キャットフード (553137)	メーカー，内容量，原産国，粗繊維，粗脂肪，粗灰分，水分，粗タンパク質

- ブランド名，商品名，メーカー名などの重要な属性が抽出対象となっていなかったので，これらを分析対象として加えた．

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．アノテーション時には，後述する 3.2.1 節の方法で作成した属性-属性値のリストを提示し，これらと類似する表現をアノテーションするよう依頼した．また．アノテーションにあたり作業者に以下の点を指示した．

長い表現をとる 「フランスのブルゴーニュ産ワインです」という文があった場合,「フランス」,「ブルゴーニュ産」をそれぞれアノテーションするのではなく,「フランスのブルゴーニュ産」をアノテーションする．

記号で区切る 「フランス・ブルゴーニュ産ワインです」という文があった場合，記号「・」で区切り，「フランス」,「ブルゴーニュ産」をそれぞれアノテーションする．ただし固有名詞（e.g.,「カベルネ・ソーヴィニョン」），数値（e.g.,「3,000 ml」），サイズ（e.g.,「19.5×24.1×8.0 cm」），数値の範囲（e.g.,「10〜15cm」）の場合は例外とし，記号があっても区切らない．

括弧の扱い 括弧の直前，中にある表現が共に属性値と見なせる場合は別々にアノテーションする．例

表 2: 分析データの規模

カテゴリ	文数	属性値数
ワイン	355	262
シャンプー	638	490
プリンターインク	286	375
T シャツ	720	357
キャットフード	382	132
合計	2,381	1,702

えば「ブルゴーニュ（フランス）のワインです.」の場合,「ブルゴーニュ」,「フランス」を個別にアノテーションする．一方,「シャルドネ（100%）」の場合は,「シャルドネ（100%）」をアノテーションする．

以上の作業により得られた分析対象データの規模を表 2 に示す.

3　商品の属性値情報抽出システム

3.1　商品データの特徴

オンラインショッピングサイト上の商品データの特徴として以下の点が挙げられる．

1. 商品カテゴリ数が多い.
2. 一部の商品ページには表や箇条書きなどの形式で整理された属性情報が含まれている.

一般にオンラインショッピングサイトの商品カテゴリ数は多く，例えば楽天であれば 4 万以上のカテゴリが存在する．そのため，それぞれのカテゴリにおいて学習データを準備することはとてもコストの高い作業となる．その一方で，一部の商品ページにおいては図 1 に挙げたように，商品の属性情報が表や箇条書きなどを使って整理されている場合がある．これら半構造化データは店舗ごとにその形式が異なるものの，いくつかのパターンを用いれば，そこから属性-属性値情報をある程度の精度で抽出することができる.

3.2　抽出システム

エラーの原因の特定を容易にするためにはシンプルなシステムが望ましく，また上述した商品データの特徴を考慮すると教師あり学習に基づく抽出手法は難しい．そこで今回は商品ページに含まれる半構造化データから属性-属性値辞書を構築し，この辞書を使った辞書マッチによって商品ページのタイトル，商品説明

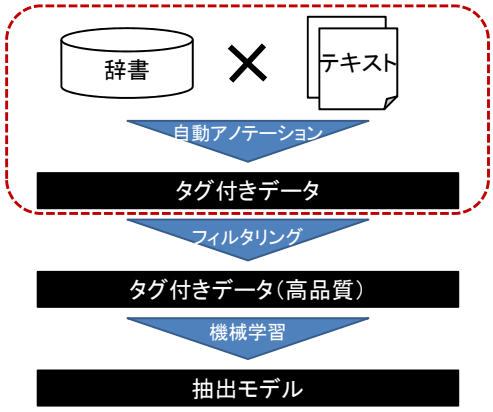


図 2: Distant supervision の流れ

文，販売方法別説明文から属性値を抽出するシンプルな情報抽出システムを実装した.

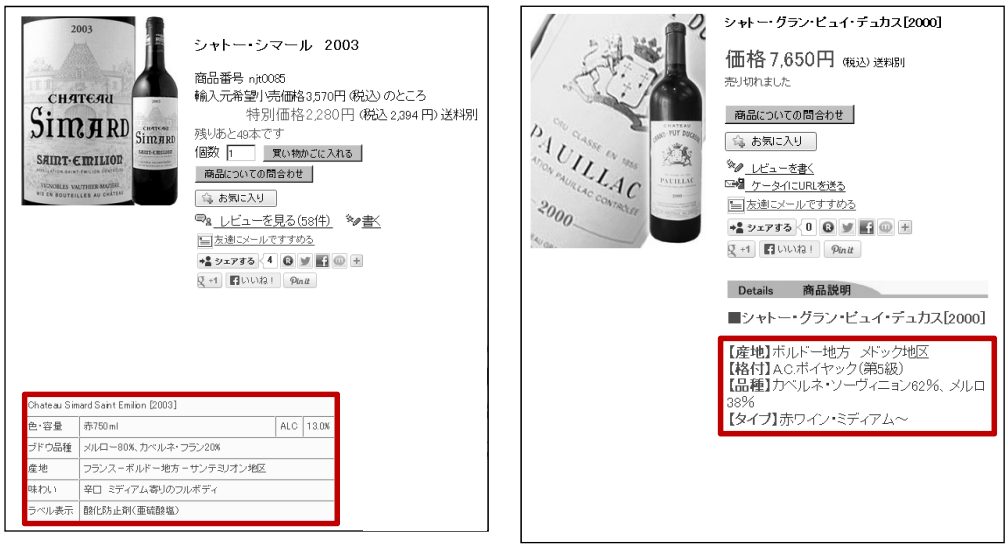
近年，図 2 に示すような Distant supervision に基づく情報抽出手法が多く提案されている [1, 5, 4, 2, 6]．これらは Freebase や Wikipedia の Infobox などの人手で整備された辞書を活用してテキストデータに対し自動でアノテーションし，これを訓練データとして抽出規則を学習する．本システムで Freebase や Wikiepdia の Infobox を用いない理由は，これら辞書にはオンラインショッピングで有用となる商品の属性-属性値が記述されていない商品カテゴリが多く，教師データを自動構築する際の辞書データとしては利用できないためである.

本手法は単純なものであるが，これは Distant supervision における初期タグ付きデータ作成部分に相当する（図中の赤破線の部分）．多くの手法では，この後，固有表現抽出と組み合わせたフィルタリングや，統計量を用いたフィルタリング等の処理を行ってタグ付きコーパスから false-positive/negative を減らすように工夫している．そのため，ここでのエラー分析の結果は商品の属性値抽出のみならず，Distant supervision に基づく一般の情報抽出タスクにおいても，どのようなエラーについて後続のフィルタリング処理で考慮しないといけないのか，を示唆する有用な知見になると考えられる.

以下，属性-属性値辞書の構築方法，および辞書に基づく属性値抽出方法について述べる.

3.2.1　属性-属性値辞書の構築

本節では属性-属性値辞書の構築方法について簡単に述べる．詳細な構築方法については，論文 [3] を参照されたい.



(a) 表　　(b) 箇条書き

図 1: 商品ページ中に含まれる半構造化データ（枠で囲まれた部分）

前述したように一部の商品ページには表や箇条書きなどの半構造化データが含まれており，今回はこれらを手がかりに辞書を構築する．まずドメイン特有の属性を得るため，正規表現パターン <TH>.+?</TH> を使って表のヘッダーから属性を獲得する (<TH> は表のヘッダーを表す HTML タグ)．獲得された属性のうち保存方法, その他, 商品説明, 広告文責, 特徴, 仕様は適切な属性と見なせないため除く．

続いて属性-属性値の組を抽出するため，以下に示す正規表現パターンを商品ページに適用し，[ANY] にマッチした表現を [ATTR] に対応する属性の値として抽出する．

P1: <T(H|D)>[ATTR]</T(H|D)><TD>[ANY]</TD>

P2: [P][ATTR][S][ANY][P]

P3: [P][ATTR][ANY][P]

P4: [ATTR][S][ANY][ATTR][S]

ここで [ATTR] は事前に獲得しておいた属性を表す文字列，[ANY] は任意の文字列，[P] は○，●，◎，□，■，・，☆，★，【，＜，[のいずれかの文字，[S] は：，／，】，＞，] のいずれかの文字を表す．なお P4 において，[ANY] は最初に出現した [ATTR] の値とする．

以上の操作を楽天市場のワイン，シャンプー，プリンターインク，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データに対して適用した．獲得された属性-属性値をカテゴリ毎に 400 件無作為に抽出し，正しい関係になっているかどうかを 1 人の被験者により評価した．獲得された属性-属性値の数，および正解率を表 3 に示す．

表 3: 属性-属性値の数と正解率

カテゴリ	属性-属性値の数	正解率 [%]
ワイン	3,940	80.8
シャンプー	6,798	83.3
プリンターインク	956	71.3
T シャツ	10,227	43.5
キャットフード	797	83.0
合計	2,381	1,702

最後にワインカテゴリに対して獲得された属性-属性値の例を表 4 に示す．

3.2.2 属性値情報の抽出

まず入力文を形態素解析し，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．この時，抽出された属性値からさらに別の属性値を取ることは考えない．また，誤抽出の影響を少なくするため属性値が数値のみからなる場合は抽出しなかった．形態素解析器には JUMAN 7.01[3] を用いた．

2 節で述べたデータに対し属性値の抽出を行った時の true-positive/false-positive/false-negative の事例数を表 5 に示す．5 カテゴリ中 4 カテゴリでは，辞書の正解率は 80%近かったにも関わらず，それらを用いて行った自動抽出の精度が低いことがわかる．次節では false-positive/negative の事例について調査する．

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

表 4: 自動構築した属性-属性値辞書の例（ワイン）

品種	容量	産地	生産者	タイプ	度数
シャルドネ	750ML	フランス	ファルネーゼ	辛口	１２％
メルロー	720ML	イタリア	マス デ モニストロル	赤	１２．５％
シラー	375ML	スペイン	ルロワ	白	１１．５％
リースリング	500ML	チリ	M. シャプティエ	フルボディ	１１％
グルナッシュ	1500ML	ボルドー	マストロベラルディーノ	やや甘口	１３％
サンジョベーゼ	360ML	シャンパーニュ	サンテロ	甘口	１３．５％
メルロ	200ML	オーストラリア	サルタレッリ	やや辛口	１４％
マカベオ	3000ML	アメリカ	カビッキオーリ	ライトボディ	１０％
テンプラリーニョ	1800ML	ドイツ	フォントディ	ミディアム	１２度
シラーズ	1000ML	アルゼンチン	カ ルガーテ	ロゼ	１４度未満

表 5: true-positive/false-positive/false-negative の数

カテゴリ	TP	FP	FN
ワイン	159	123	100
シャンプー	335	337	199
プリンターインク	171	67	198
T シャツ	95	446	316
キャットフード	111	84	18
合計	871	1,057	831

4 エラー分析

4.1 False-positive の分析

false-positive となった 1,057 事例について，以下の項目を順次チェックすることで分類を試みた．

1. 正しい属性-属性値に基づいて属性値が抽出されたかどうか
2. 属性値を抽出するべき商品ページかどうか
3. 商品に関係するパッセージから属性値が抽出されたかどうか

以上のチェック項目をパスした抽出結果は，適切な商品ページの適切なパッセージから適切な属性-属性値に基づいて抽出されたものであるにも関わらず誤抽出と判断されたものである．そこで，最後に残った事例をさらに分類した．

本節では分類結果および各事例について述べる．

4.1.1 正しい属性-属性値に基づいて属性値が抽出されたかどうか

属性値の抽出は商品ページの半構造化データより自動構築した辞書に基づいて行っている．辞書は自動構築しているため，誤った属性-属性値の組も含まれている．そこでまず，誤った属性-属性値に基づいて抽出された結果であるかどうかを確認した．この項目に該当する事例数は 712 であり，false-positive 事例の 67.4%に相当する．これより質の高い辞書を構築することが本タスクにおいては重要であることがわかる．

この項目に該当する事例を減らすためには，辞書構築の方法を見直す必要がある．今回は表・箇条書きデータに注目して辞書を構築しているため，商品ページ中のこれらデータの解釈をより正確に行う必要があることを意味している．また，表・箇条書き以外の手がかり（例えば語彙統語パターン）も取り入れることで辞書構築の性能の向上が見込める．

4.1.2 属性値を抽出するべき商品ページかどうか

楽天では商品ページを商品カテゴリに登録する作業は店舗によって行われており，そこには誤りも含まれている．そこで誤って分析対象カテゴリに登録された商品ページかどうかを確認した．誤ったカテゴリに登録されている商品ページは今回のデータセット中では 4 件[4] あり，そこに含まれる false-positive 事例数は 53 件 (5.0%) であった．

このような誤りを除くためには，与えられた商品ページがカテゴリに該当するものであるかどうかを判定する処理が必要である．例えば，村上ら [7] は辞書に基づく方法で商品ページが正しい商品カテゴリに登録されているかを判定する手法を提案している．このような手法を用いることで，この項目に該当する事例を減らすことができると考えられる．

4.1.3 商品に関係するパッセージから属性値が抽出されたかどうか

商品ページには当該ページで販売している商品以外の商品について記述されることも多い．例えば，商品ページ閲覧者を店舗サイト内で回遊させるために，当該ページで販売されている商品以外の商品の広告を

[4] シャンプーカテゴリにシャンプーの容器，化粧品，T シャツカテゴリに帽子，キーホルダーが登録されていた．

掲載していたり，検索結果に頻繁に表示されるようにキーワードスタッフィングなどの操作が行われている商品ページがある．そこで 3 番目の項目として，当該ページにて販売されている商品に関係するパッセージから属性値抽出が行われたかどうかを確認した．結果，99 件 (9.4%) の事例がこの項目に該当した．このうち 90 件は，以下のような他の商品へのナビゲーションであった．

- 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♪※順次追加中！(ワイン)
- ミルボンメーカー 一覧はこちら（シャンプー）
- 色違い”ナチュラル色” ’’ＴＨＥＲＥ ＡＲＥ ＷＡＶＥＳ ＮＡＴ’’クルーネックＴシャツ (T シャツ)
- 《チャオ缶 国産原産国》(キャットフード)

同一店舗サイト内の他のページへのリンクの有無や店舗ごとの商品ページのテンプレートを認識した結果を利用することで，このような事例は減らせるのではないかと考えられる．

残りの 9 件はキーワードスタッフィングが行われた領域から抽出されたものであり，キーワードスタッフィングの検出を前処理として用いることで，これらの事例を削除することが期待できる．

4.1.4 残りの誤り事例はどんなものか？

ここまでの項目をパスした抽出結果は，適切な商品ページの適切なパッセージから適切な属性-属性値に基づいて抽出されたものであるが誤っている．このような誤りは 200 件 (18.9%) あり，これら事例をさらに分類すると以下のようになった．

人手アノテーションと部分一致 人手アノテーションと部分一致している事例が 84 件あった．このうち以下の例のように正解とみなしても問題ない事例が 37 件あった．(波線が人手アノテーション，直線が自動抽出結果)

- ドメーヌ・レ・グリフェはボジョレー産地 の南 産地 に位置する歴史あるドメーヌです。
- 国内製造国 製 製造国 ヘアケア品
- 薄手のコットン素材 素材 で着心地抜群。
- 表記Ｌ サイズ

残りの 47 件中 41 件はシャンプーの成分に関するものであり，以下の例のように人手アノテーションと部分一致しているものの，これが抽出されても意味をなさないものであった．

- ２－アルキル－Ｎ－カルボキシメチルヒドロキシエチルイミダゾリニウムベタイン、ラウロイルメチル－Ｂ－アラニン成分 ＮＡ液成分、ヤシ油脂肪酸アミドプロピルベタイン液

このような事例は属性-属性値辞書のカバレージを改善することで減らせると考えられる．

他のエンティティの部分文字列からの抽出 次に多かった誤りは他のエンティティの部分文字列から抽出している事例であり 39 件あった．これらはエンティティのタイプから組織名やイベント名，型番，ブランド名などの固有表現，ドメイン固有の用語，一般的な名詞句に分類できた．

固有表現の一部から抽出されていた例を示す．(太字がエンティティ)

- **フランス** 産地 **革命**の戦いの舞台にもなった歴史あるシャトー。
- 醸造方法も**シャトー・マルゴー** 産地 と 同じ手法をとって、 セカンドながらも品質は他の特級シャトーに匹敵するほどです。
- ２０１１年度の**トロフィー・リヨン・ボジョレー** 産地 **・ヌーヴォーコンクール**では見事金賞を受賞！
- １円３個まで リピート歓迎 ＣＡＮＯＮ（キヤノン）対応の純正互換インクカートリッジＢＣＩ－６ＰＭ（残量表示機能付）(関連商品 **ＢＣＩ－６ＢＫ** カラー ＢＣＩ－６Ｃ ＢＣＩ－６Ｍ ＢＣＩ－６Ｙ ＢＣＩ－６ＰＣ ＢＣＩ－６ＰＭ ＢＣＩ－６Ｒ ＢＣＩ－６Ｇ)
- アメリカンイーグル（ＡＭＥＲＩＣＡＮ ＥＡＧＬＥ）は、**ＡＢＥＲＣＲＯＭＢＩＥ** ブランド **＆ ＦＩＴＣＨ**（アバクロンビー＆フィッチ）と並んで人気のカジュアルブランドで、北米では８００店舗の直営店を持っています。
- **ブラック** 色 **メタル** ページ正規ライセンスＴシャツ販売

このような事例は全部で 24 件あり，前処理として固有表現認識を行い，固有表現の一部からは属性値を抽出しない，等のルールを適用することで事例を減らす

ことができると考えられる．ただ，ブランド名等は従来の固有表現タイプではカバーされていないため，従来にはないタイプの固有表現の認識技術が求められる．

次にドメイン固有の用語から抽出していた例を示す．

- また、フランスで最も古いＡＯＣ、**ブランケット・ド・リムー** 産地 を産出します。
- **ボジョレー** 産地 **・ヌーヴォー**２０１３年（新酒）！
- パーマ・デジパー（デジタルパーマ）・縮毛矯正ストレートパーマ　エアウエーブ**・水** 成分 **パーマ・**フィルムパーマなどパーマのウエーブを長持ちさせたい方に。

このような例は全部で 12 件であった．固有表現の場合と同様に，ドメイン毎に専門性の高い用語を抽出するなどし，用語の部分文字列からは属性値を抽出しないなどのルールを設ける必要があると考えられる．

最後に名詞句の一部から抽出されていた事例を示す．このような事例は以下の 3 件であった．

- ２位にルイ・ロデレール・ブリュット、７位にテタンジュ・ブリュットなど　大手の**シャンパン** タイプ **ハウス**も名を連ねています。
- ジョエル・ファルメ氏が引き継いだころは、栽培した葡萄を**シャンパン** タイプ **メーカー**に売っていましたが、現在は、葡萄の栽培・醸造・瓶詰めまで行うＲＭ（レコルタン・マニピュラン）です。
- **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。

これらを除くためには名詞句の構造を解析し，主辞以外の部分からは属性値を抽出しない，等の処理が考えられる．

当該商品の属性値の説明とは関係ない記述からの抽出
このような事例は 37 件あった．以下に例を示す．

- スペイン 産地 のロマネ・コンティで知られるヴェガ・シシリア社がハンガリーで造るワイン。
- 米国ではアメリカンイーグル、アバクロ、ＧＡＰ ブランド（ギャップ）は３大アメカジブランドとして、３つとも同じくらいの知名度となっています。
- Ｍ，Ｌ　モデル着用サイズ：Ｍ（モデル　身長：１７０ＣＭ，　体重：５８ＫＧ，　ウエスト：７２ＣＭ，　ヒップ：９０ＣＭ，　胸囲：８８ＣＭ，肩幅：４４ＣＭ 肩幅，　首周り：３７ＣＭ）
- 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。
- レビューで５％ 粗脂肪 ＯＦＦクーポン！

上の例からわかるように，ワインはワイナリーに関する記述から，Ｔシャツはブランドの説明およびモデルの体型に関する記述から，キャットフードはその利用方法やクーポンに関する記述から誤った情報が抽出されており，カテゴリによってばらつきがある．そのため，カテゴリ毎に商品ページ内の各文が何について言及しているのか，といったタイプを識別する処理が必要になると考えられる．

属性値の多義性に起因する誤抽出　このような事例は 33 件あった．この中で最も多かったタイプはサイズに関する属性値であり，16 件であった．以下に例を示す．

- 着丈５９ＣＭ、身幅４２ＣＭ 肩幅、袖幅１７ＣＭ
- ５４．５ＣＭ 身幅
- ＥＭＰＯＲＩＯ　ＡＲＭＡＮＩ　２０１３ＳＳ サイズ 新作　半袖Ｔシャツ　Ｌ１Ｔ１５Ｊ　Ｌ１Ｑ４Ｊ　１００　ホワイト　エンポリオアルマーニ　ＥＡ　クルーネック

1 つ目の例のように，サイズに関する情報は属性名とともに属性値が記述されることが多いため，属性名に相当する表現と属性値がどのくらい離れた場所に記述されているか，という指標を考慮することで誤りを減らせる可能性がある．2 つ目の例は表のセルに記述されたものであった．そのため，表形式で記述されたデータの理解も重要な処理と考えられる．3 つ目の事例は，春夏シーズンを意味するの SS（Spring Summer）とサイズを誤っている例である．このような事例を除くには従来から研究されているような多義性解消技術の導入が必要である．

次に多かったタイプは割合に関する表現であった．このタイプの事例は 9 件であった．以下にその例を示す．

- ピノノワール７０％、ピノムニエ２０％ 度数、シャルドネ１０％
- 粗たん白質：４．０％以上 粗脂肪、粗脂肪：０．１％以上、粗繊維：０．１％以下、粗灰分：１．０％以下、水分：９４．０％以下、エネルギー：約１５ＫＣＡＬ／袋

- ０．０５%以上 粗脂肪

1つ目の例のように混合比が素材と一緒に併記されることが多い．そのため，素材に相当する表現の間に挟まれる割合表現を抽出対象としないことでエラーを減らせると考えられる．またサイズ同様，属性名にあたる表現と併記されることがあるため，属性名との距離を考慮することである程度事例数を減らすことが期待できる．また割合も表形式のデータで記述されることがあるため，表データの理解は重要であろう．

以下は本来であれば，ワインの属性「タイプ」の値として抽出されるべきであるが，地名として抽出されてしまった例である．

- ＮＹタイムズで、ベストシャンパーニュ産地（４０ドル以下）に選ばれました。
- モエ・シャンドン・ドンペリニヨンの最高級品、通称「ドンペリ・ゴールド」最高の葡萄を熟成させ生産量が極めて少なく本場フランスと日本でしか手に入れることのできない究極の「幻のシャンパーニュ産地」と呼ばれています。

上の例は「40 ドル以下」，下の例は「フランスと日本でしか手に入れることのできない」という表現から「シャンパーニュ」が「地名」ではなく「タイプ」の意味で使われていることが（人間には）わかる．しかしながら，この処理を計算機で実現することは現時点では難しいと考えられる．

メタファーに起因する誤抽出　このタイプに該当する事例は 5 件あり，すべて「ボジョレー」に関するものであった．以下に例を示す．

- 本物のボジョレー 産地 の味わいを　感じさせてくれる、自然派！
- ボジョレー 産地 に求める要素をすべて備えていると言っても過言ではありません。

「ボジョレー」はワインの産地の 1 つであるが，ここでは産地としてではなく,「ボジョレー産のワイン」という意味で用いられている．上の例は「の味わい」という表現に注目することで「産地」でないことがわかる．一方下の例は文単体では「産地」という理解も可能である．しかしながら，当該文の直前の文が「彼らのスタイルは飲み心地が良く、フルーティで果実味が豊か。」であることを考えると「産地」ではないことがわかる．このように，情報抽出システムの性能の向上には文を跨いだ言語処理が必要になる．

形態素解析の過分割による誤抽出　形態素解析により過分割されたために誤って抽出された事例が 1 件あった．以下に示す．

- トカイ・フルミント・ドライ・マンデュラス 品種［２００６］（オレムス）

マンデュラス (mandulas) とはハンガリー語でアーモンドを意味する語である．形態素解析器の辞書にマンデュラスが登録されていなかったため，過分割されてしまい誤った属性値が抽出されていた．マンデュラスのような語が形態素解析器の辞書にあらかじめ登録されていることは期待できないため，あるドメインに関するテキスト集合から自動的に語彙を獲得し，形態素解析器の辞書を動的に拡充する手法が必要であると考えられる．

商品ページ内の誤った情報からの抽出　誤った情報が商品ページに記述されており，そこから誤った属性値が抽出されている事例が 1 件あった．以下に示す．

- アリミノ　ミントシャンプー　フローズンクール ２２０ＭＬ 容量

商品タイトルに 1000ml と記述されており，商品画像も 1000ml のものであったことから 220ml は誤りであることがわかった．このように抽出元となるテキストの信頼度や，画像データなどのテキスト以外の情報を考慮することも精度の向上に必要である．

4.2 False-negative の分析

false-negative に該当する事例は全部で 831 件あった．分析にあたり，キャットフードカテゴリについては全 18 件，キャットフード以外のカテゴリからは無作為に 50 件ずつ選び出した．そして，以下の条件のいずれかに一致する事例を削除して残った 188 件について分析を行った．

- 誤ったカテゴリに登録された商品ページ．
- 人手アノテーションと部分一致し，かつ正解と見なしても問題ないもの．

4.2.1 異表記が辞書に含まれていない

異表記が辞書に含まれていないため false-negative となっている事例が 100 件（53.2%）あった．表 6 に異表記が辞書に含まれていない属性値のタイプと例を

示す．組織名，地名，割合表現，人名など既存の固有表現のタイプが見てとれる．そのため，固有表現のタイプと属性の間に変換ルールを設けることで，辞書に含まれていない属性値についても固有表現認識により抽出できる可能性がある．しかしながら，この操作によって false-positive の数が増えてしまう可能性があることに留意する必要がある．

4.2.2 異表記が辞書に含まれている

属性値自身は辞書に含まれていないが，その異表記が辞書に含まれている事例は 69 件 (36.7%) あった．異表記のタイプ，各タイプの数および例を表 7 に示す．空白，中黒，ハイフンの有無, 入れ替わり，長音とハイフンの入れ替わり，接辞の有無，翻字の違い，小数点の扱い，送り仮名の有無など，テキスト中と辞書中の表現の編集距離などの類似度を考慮した柔軟なマッチングを行うことで改善できる事例が多いことがわかる．一方で，略語，翻訳，言い換えなど，事前の知識獲得を必要とする事例も見られる．

4.2.3 抽出手法の問題

辞書に正しい属性-属性値の組が登録されているにも関わらず，手法の問題により抽出されなかった事例が 19 件（10.1%）あった．この中で最も多かったタイプは数値単体からなる属性値であった（13 件）．今回は，誤抽出の影響を減らすため数値のみの属性値は抽出しないようにしており，これが原因で数値のみの属性値が抽出されなくなっていた．数値に関する抽出手法を洗練することで，このタイプの誤りは減らせると考えられる．

残り 6 件のうち 3 件は，辞書エントリとテキストの最長一致による属性値抽出方法が問題となっていた．具体的には，正解の属性値よりも文字列長の長い誤った属性値が先に抽出されてしまい，正しい属性値が抽出されなくなっていた．文字列長だけではなく，辞書エントリとしての正しさも考慮にいれて抽出を行うことで改善できる可能性がある．

残りの 3 件は属性値に多義性がある場合であった．属性値抽出を行う際，店舗頻度 [5] を元に多義性解消を行っているが，この処理が誤っていた．この事例については，より正確な多義性解消を導入することで改善される見込みがある．

[5] 論文 [3] を参考にされたい．

5 おわりに

本稿では Project Next NLP 情報抽出タスクについて述べた．具体的なタスクとして商品説明文からの商品属性値の抽出を設定し，属性-属性値辞書に基づく情報抽出システムを実装した．そしてこのシステムを楽天市場のワイン，シャンプー，プリンターインク，T シャツ，キャットフードカテゴリに登録された商品データに対して適用し，false-positive / negtive にどのようなタイプのエラーが存在するのか調査した．

エラーの削減に必要なデータおよび処理としては以下が考えられる．

- 質とカバレージの高い辞書
- 詳細な固有表現タイプの認識技術
- 属性値抽出対象箇所の同定
- 属性値を抽出する際の多義性解消
- 知識獲得（新規辞書エントリの獲得, 辞書エントリの同義語獲得）
- 辞書エントリとテキスト中の表現の柔軟なマッチング

本タスクは Distant supervision におけるタグ付きデータ作成方法とも見なせる．今後は上記の問題点を考慮したより高品質なタグ付きコーパス作成方法を実装し，それを基に機械学習ベースの属性値情報抽出システムを開発する予定である．

参考文献

[1] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 1003–1011, 2009.

[2] Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. Modeling missing data in distant supervision for information extraction. In Transactions of the Association of Computational Linguistics – Volume 1, pp. 367–378, 2013.

表 6: 異表記が辞書に含まれていない属性値の例

タイプ	数	例
型番	27	(適応機種, ＰＸ－４３４Ａ), (適応機種, ＯＦＦＩＣＥＪＥＴ　ＰＲＯ　Ｋ５４００)
組織名	15	(生産者, ドメーヌ　レ　グリフェ), (生産者, ＣＨＡＴＥＡＵ　Ｄ’ＹＱＵＥＭ), (メーカー, わんわん), (メーカー, 日本ヒルズ・コルゲート)
商品名	14	(商品名, ナカノ　デスピナ　シャンプー　スキャルプ　ボリュームダウン　８００ＭＬ)
ブランド名	10	(ブランド, ＶＡＮＩＬＬＡＦＵＤＧＥ), (ブランド, ポスターリスト)
割合	7	(粗繊維, ０．１%), (粗タンパク質, １１．３%以上)
素材	7	(成分, （ジヒドロキシメチルシリルプロポキシ）ヒドロキシプロピル加水分解ケラチン), (成分, 和漢植物エキス)
サイズ	6	(サイズ, ５００サイズ用), (サイズ, 着丈６５ＣＭ、身幅４９ＣＭ、袖幅１９ＣＭ)
人名	6	(生産者, ジョエル・ファルメ), (生産者, ピエール・デュルディリ)
容量	4	(容量, １８．２ＭＬ), (内容量, ５７Ｇ×１２カップ)
地名	4	(産地, コート・ドパール), (産地, 島根県)

表 7: テキスト中の表現と辞書エントリの表記の違い

タイプ	数	例	
		テキスト	辞書
空白，中黒，ハイフンの有無, 入れ替わり	18	ＰＩＸＵＳ　９９０Ｉ	ＰＩＸＵＳ９９０Ｉ
長音とハイフンの入れ替わり	11	オレスー８リン酸ＮＡ	オレスー８リン酸ＮＡ
略語	13	ＸＬ デビフ １００%　ＣＯＴＴＯＮ	ＥＸＴＲＡ　ＬＡＲＧＥ デビフペット Ｃ１００%
翻訳	9	ＳＡＵＴＥＲＮＥＳ ホリスター ラウレス硫酸ＮＡ	ソーテルヌ ＨＯＬＬＩＳＴＥＲ ラウレス硫酸ナトリウム
言い換え	7	１４．５% ヴーヴ・クリコ社	１４．５度 ヴーヴ　クリコ　ポンサルダン
接辞の有無	4	ペンフォールド	ペンフォールド社
翻字の違い	2	ラウロイルジモニウムヒドロキシプロピル加水分解ケラチン	ラウリルジモニウムヒドロキシプロピル加水分解ケラチン
小数点の扱い	1	２%以下	２．０%以下
送り仮名の有無	1	８０Ｇ×４８缶入	８０Ｇ×４８缶入り
その他	3	—	—

[3] Keiji Shinzato and Satoshi Sekine. Unsupervised extraction of attributes and their values from product description. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pp. 1339–1347, 2013.

[4] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 721–729, 2012.

[5] Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 118–127, 2010.

[6] Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. Filling knowledge base gaps for distant supervision of relation extraction. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 665–670, 2013.

[7] 村上浩司, 関根聡. カテゴリに強く関連する語の発見と商品データクリーニングへの適用. 言語処理学会 第 18 回年次大会 発表論文集, pp. 195–198, 2012.