ネットオークションにおける属性検索のための出品情報文書からの属性抽出

西村 純† 宮崎 林太郎† 前田 直人† 森 辰則†

翁 松齢‡ 石川 雄介‡ 小林 寛之‡ 田中 裕也‡

†横浜国立大学大学院環境情報学府 ‡ヤフー株式会社

E-mail: †{jun-n, rintaro, n-maeda11, mori}@forest.eis.ynu.ac.jp

‡{shou, yuishika, hkobayas, yuutanak}@yahoo-corp.jp

1. はじめに

近年，インターネットなどの普及によって Web を介して何かを探す，誰かとやり取りを行なう，情報を集めるなど, Web 上での作業が生活の中で必要不可欠となってきている．その中の 1 つとして，最近盛んに行なわれているのが「ネットオークション」である．利用者はどこからでも気軽に参加でき，「売りたい」や「買いたい」という気持ちを満たすための重要な手段となっている．

現在のネットオークションでは，図 1 のように，サイズや色などの属性情報を検索語として入力した場合, 本当に欲しい属性記述のある出品情報以外に, 属性情報以外に現れる文字列に一致する出品情報も検索されてしまうという問題がある．本研究では，ネットオークションの出品情報文書に多数存在する商品の属性情報に着目し, それらの情報を機械学習に基づき，高い精度で抽出し，抽出された属性, 属性値の対の同定を行うことによって利用者が望むような柔軟な検索の実現を目的としている．

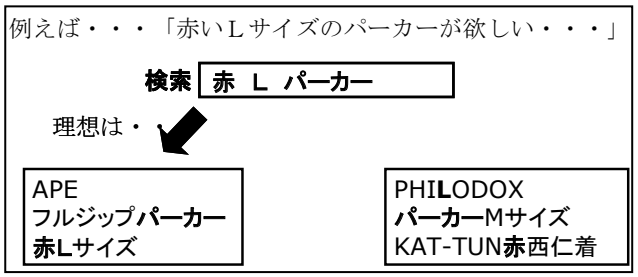

図 1：現在のオークション検索の問題点

2. 先行研究

テキストからの必要な情報を抽出する研究として, 中野ら[1]は日本語固有表現抽出において，文節区切りを行ない, 文節内の情報を素性としてチャンカーに与えることを提案し, 各文節の長さに応じて素性展開を行なうことによって, 文脈長を固定したモデルでは用いることのできなかった情報をチャンキングに利用している．

一方，森ら[2]は動向情報編纂のためのテキストからの統計量の自動抽出として, 統計量名を構成する表現が何であるかを検討し, その構成要素を種別ごとに区別して抽出することを目的としている．

また，構成要素の対の同定に関する研究として飯田ら[3]が意見抽出を目的として機械学習に基づく手法を用いて属性と評価値対を同定する方法を提案している．

本研究では, ネットオークションに関する属性情報の抽出を固有表現抽出などでも用いられている一般的な手法によって行なっている．また，一部のカテゴリに特化しない自動抽出の実現のために角川類語新辞典[4]からの分類情報を利用し, 表層表現に依存しなくてもある程度の精度が得られることを実験により示している．また，構成要素の抽出処理で抽出された属性, 属性値の対の同定に関する実験を単純な手法を用いて行い, 後に述べるように高い精度が得られることを確認している．

3. 基本的なアプローチ

本研究で提案する出品情報文書を対象とした属性情報の自動抽出処理は，大きく 2 つの処理に分けられる．1 つ目は，図 2 に示すような構成要素の抽出の処理である．この処理はさらに 2 つの処理から構成される．すなわち，属性情報に対する注釈の付与による学習用コーパスの構築，素性展開，機械学習，自動抽出器の作成という学習フェーズと自動抽出器を利用した未知の出品情報文書からの属性，属性値の抽出を行う抽出フェーズである．

また，2 つ目の大きな処理として，1 つ目の処理で抽出された構成要素の対の同定を判定する処理である．ここでは，飯田ら[3]が先行研究においてトーナメントモデルと比較するモデルとして用いた, 照応解析の単純なモデルである Soon らの解析モデル[5]を属性情報の対の同定の判定に適用している．

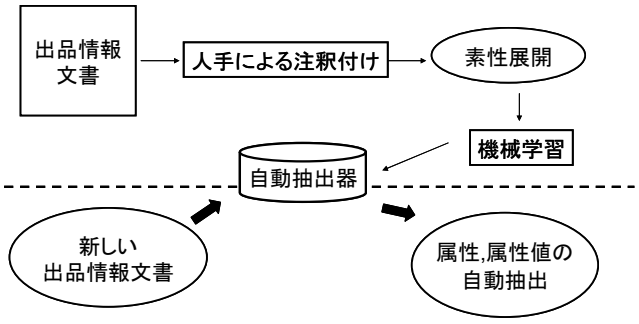

図 2：構成要素の抽出のアプローチ方法

表 1:本研究で定義した属性情報

属性名	属性値の例
色	黄色，イエローなど
素材	ポリエステル 50%，綿 100%など
サイズ	着丈：65cm，M サイズなど
形状	半袖，ノースリーブなど
状態	新品，未使用，古着など
定価	定価 2000 円など
製造場所	日本製，made in USA など
シーズン/モデル	秋冬モデル，1970 年代など
デザイン	花柄，ストライプなど
その他	重さなどあまり出てこないもの

3.1. 人手による注釈付けによる学習用コーパスの構築

本研究における抽出対象は属性と属性値の組である. 属性とは商品の様態を表わす観点に対応する表現であり, 属性値とは属性に対する様態の内容を示す表現である. 一般の属性・属性値抽出においては＜事物，属性，属性値＞の 3 項組を抽出するが，オークションの出品情報文書においては，1 文書につき 1 つの商品(事物)について記述されており，3 項組のうち事物の部分はその商品に決まるので，ここでは抽出対象としない.

3.1.1. 注釈付けを行う属性，属性値

注釈付けを行う属性，属性値としては，表 1 に示すものを対象としている．これらの情報を選んだ理由としては，教師情報となるコーパスを作成する際の注釈者間の判断の揺れを少なくすること, 利用者が検索の対象として必要だと感じていることがあげられる．なお，現段階の研究においては, 属性情報による検索の需要が最も高いと考えられるファッションのカテゴリについて検討を行なっている．上記の 10 項目において，該当オークションの説明に関する情報すべてについて XML タグを付与することにより注釈付けを行う．属性には<attr id=”a○○○_△△”>タグを，属性値には<val id=”v○○○_△△”>タグを付与していて，先頭一文字が a もしくは v であって、後接する文字列が共通である id が付与されている．出品情報文書に対する注釈付けの例を図 3 に示す.

3.1.2. 出品情報文書中の表現への注釈付け方法

出品情報文書中の属性, 属性値の現れ方には幾つかの場合が考えられる．利用者にとって，検索の際にどのように指定できるかを検討した上で, 以下のように注釈づけを行う．

① 属性，属性値が組として現れない場合は単独でも注釈付けを行う．
② 属性と属性値が一つの複合語になっている場合は，分解をし，個別に注釈付けを行う．“属性値－属性”の順で現れることが多い．複合語をまとめて 1 つの属性値と考えることもできるが表現に属性の情報が現れているのでそれを組の情報として生かす意味で上記のように注釈付けをする．
③ 属性が階層構造を持つ場合には，階層を考慮せずに，個別に注釈付けを行う．

::AID::
53916975
::TITLE::
一撃落札！！<val id=“v001_01”>古着</val>シャツ！！
<val id=“v001_02”>黄</val> <attr id=“a001_02”>色</attr>
<val id=“v001_03”>M</val> ボーリング
::DESCRIPTION::
<val id=“v001_04”>新品</val>！！<attr id=“a001_05”>サイズ</attr>は、
<attr id=“a001_06”>着丈</attr>：<val id=“v001_06”>69</val>、
<attr id=“a001_07”>身幅</attr>：<val id=“v001_07”>49</val>、
<attr id=“a001_08”>袖丈</attr>：<val id=“v001_08”>25</val>、
<attr id=“a001_09”>肩幅</attr>：<val id=“v001_09”>43cm</val>くらいです。
::CATEGORYID::
2084030337

図 3：出品情報文書に対する注釈付けの例

3.2. 素性展開(文字単位)

本研究では, 文字を単位とする分類問題として定義された系列ラベリングに基づくチャンキングにより情報抽出を行うことを考える．そのために，出品情報文書を文字の単位に分け，各々に 6 つの素性を与えている．なお，4 章で述べる実験においては, どの素性が有効であるか検討をするために使用する素性を変えて幾つかの実験を行っている. 与えた素性は具体的には, 表層文字, 文字種, 品詞, 文節内素性，複合名詞主辞素性(以下，主辞素性と記載), シソーラス上の概念分類番号(以下，分類番号と記載)である.

文節内素性とは文節内に固有表現が存在すれば, 最も先頭に近い固有名詞の品詞細分類を, 固有名詞がなければ文節の先頭の単語を素性として用いるものである. 複合名詞主辞素性とは，連続する名詞が存在する場合，連続する名詞の最後の名詞を素性とするものである.

分類番号とは, 角川類語新辞典において各単語に付与されている番号のことである. 角川類語新辞典の語彙分類構造は十進分類になっていて，まず大項目が「自然･性状･変動･行動･心情･人物･性向･社会･学芸･物品」に大別されている．ついでこれが，それぞれ 10 個ずつの中項目に分かれている．さらに，これらが十進ずつの小項目に分けられていて, これら 3 階層における各項目番号を順番に連結してできる 3 桁の数字が分類番号である．例えば，「紫」，「赤」，「グリーン」，「カラー」など「色」に関する単語には「143」という分類番号が付与される．分類番号を素性として用いるときには，同じ範疇に属する，意味的に近い単語には同じ分類番号が付与されるので, 表層表現が異なっていても同じ素性を持つ事例として考慮される.

3.3. チャンクの表現方法

チャンキングを行なう際, チャンクの状態をどのように表現するかであるが，各種先行研究においては，各トークンにチャンクの状態を示すチャンクタグを付与する方法が利用されている．チャンクタグは，対応するトークンのチャンク内での位置を表す記号と, チャンクの種類をハイフンで結んだもので表される．本研究で用いた，チャンクの符号化手法の一つである IOE2 法では，チャンクの最終トークンに E という記号を付与し，それ以前のトークン

に記号 I を付与する. 要素以外のトークンには O が付与される．また，チャンキングを行う文字の前後 2 文字ずつ計 5 文字を文脈長とした．要素の抽出規則の学習は，図 4 の枠内の素性から対応するチャンクタグを得るような分類器を, 学習事例と機械学習手法を用いて構成することに相当する．一方，未知の文における抽出の際には，各文字ごとに枠内の素性集合を導出し, その素性集合を分類器に与えることによりチャンクタグを文末から文頭に向けて順次推定する．

位置	文字	文字種	品詞	文節内素性	主辞素性	分類番号	タグ
	素	KANJI	B-名詞-普通名詞	素材	素材	805	
i+2	材	KANJI	E-名詞-普通名詞	素材	素材	805	
i+1	は	HIRAG	S-助詞-副助詞	*	*	*	
i	レ	KATAK	B-名詞-普通名詞	レーヨン	レーヨン	907	I-val
i-1	ー	OTHER	I-名詞-普通名詞	レーヨン	レーヨン	907	I-val
i-2	ヨ	KATAK	I-名詞-普通名詞	レーヨン	レーヨン	907	I-val
	ン	KATAK	E-名詞-普通名詞	レーヨン	レーヨン	907	E-val
	。	OTHER	S-特殊-句点	*	*	*	O

図 4 : 素性集合に対するチャンクタグの推定

3.4. 構成要素の対の同定の手法

ここでは, 構成要素の対の同定を行う際の手法について詳しく述べる．まず，図 5 の(a)に示されるように，仮に学習データにおいて属性値 V に対して 4 つの属性候補が現れたとする．その 4 つの属性候補に加えて，対となる属性が存在しない場合の 5 つの組み合わせに対し, 正しい(正例)か否(負例)かの教師情報を, 人手で注釈付けしたタグの情報から構築する. これらから正例と負例の情報を学習する．そして，新しい文書に対して対の同定の問題を解く際に, 訓練時と同様に 1 つの属性値に対して一定の範囲から探索された属性候補の各々が対となる場合と, 対となる属性存在しない場合のいずれになるかを判定していくことになる．また，1 つの属性値に対して複数の属性の出現が考えられるが, ネットオークションの出品情報文書の場合, 1 つの属性値に対応する複数の属性は同一のものである場合が圧倒的に多いためいずれかの属性と対の同定ができれば正解とする．学習器には SVM light を用い，分離平面からの距離を利用して 1 つの属性値に対して正解となる属性を 1 つに決定している．

4. 実験および考察

本稿では，実験として 2 つのことを行った．

1 つ目は情報抽出において比較的標準的な手法であるチャンキング手法を用いることによって, 出品情報文書から属性, 属性値がどれくらいの精度で抽出できるかを調べるための実験である．チャンキングには SVM に基づく汎用チャンカーである YamCha を使用した. 評価に際しては, 出品情報を単位とした 5 分割交差検定を行い, それらの平均の適合率，再現率を求めた．2 つ目は 3.4.節で述べた単純な手法を用いて, どのくらいの精度で属性と属性値の対が同定できるかを調べるための実験である. 出品情報においては対となる属性と属性値が比較的近くにあると予想できるため，1 つの属性値に対し属性候補を探索する範囲を変化させて実験を行った．属性候補を探索する範囲は，出品情報文書を属性，属性値，その他の部分に分割し配列に入れ, 対象となる属性値の前後いくつの配列要素を属性候補の探索範囲とするかという問題設定にしている．

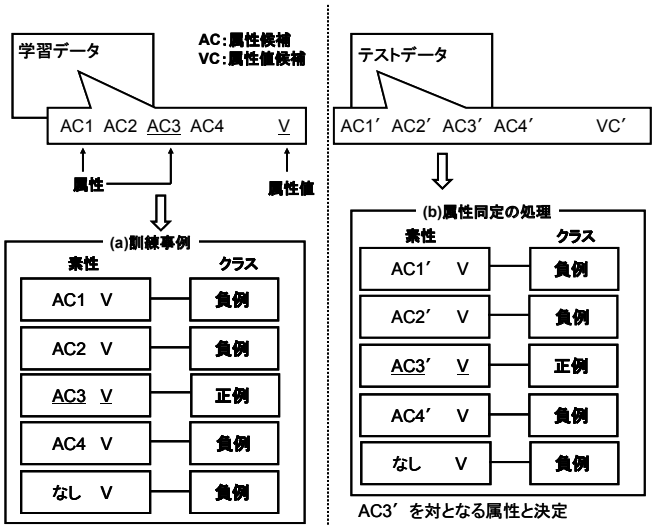

図 5 : 構成要素の対の同定のアプローチ方法

本稿では，1 商品ページすべて，前後 8 配列要素，前後 4 配列要素，前後 2 配列要素，前後 1 配列要素の 5 段階で検討している. 対の同定においても出品情報を単位とした 5 分割交差検定を行っている．

4.1. 実験データ

実験には, Yahoo!オークションに出品された商品の出品情報のうちファッションカテゴリのものを用いた. この際, 出品者に固有の記述様式による影響を排除するために, 出品者が重複した出品情報文書は用いないように考慮した. 用いたデータの詳細を表 2 に示す.

表 2 : ネットオークション出品情報文書の詳細

データ	データ数	総数		異なり数	
		属性	属性値	属性	属性値
アパレル(男性用)-トップス-シャツ-半袖	150	1422	1794	149	512
アパレル(女性用)-トップス-キャミソール	150	723	1245	91	381
アパレル(女性用)-和服-浴衣	149	928	1143	115	391

4.2. 同一出品情報文書内における構成要素の抽出精度

本節では, 学習データとテストデータに同一の出品情報文書を用いて抽出の実験を行った. 用いる素性は以下のように変化させている．

① 表層文字,文字種,品詞,文節内素性,主辞素性
② 表層文字,文字種,品詞,文節内素性,主辞素性,分類番号
③ 文字種,品詞
④ 文字種,品詞,分類番号

結果を表 3 に示す.

①と②の比較からわかるように, 分類番号を素性として与えることによって適合率の低下を招くことなく再現率が上昇していることが確認できる．しかし，表層表現に関係する素性への依存が高いために分類番号の効果は大きくないことがわかる．

③と④の比較では, 分類番号が全体的な精度の上昇に非常に有効に働いているといえる．つまり，表層表現に依存しない素性だけでもある程度の抽出精度を保てているということから, 学習用の注釈付きコーパスに現れない新しい属性，属性値であっても，既存のシソーラスに現れる表

現であれば，精度の低下を招かずに属性，属性値の抽出が行えることが期待される．特に，新しい分野の出品情報における属性，属性値の抽出において有効であると考えられる．しかしながら，表層表現を用いた場合と比較すると，抽出精度の低下が見られるので，さらなる検討を要する．

表 3：構成要素の抽出結果(トップス-シャツ-半袖)

	属性		属性値	
	適合率	再現率	適合率	再現率
①	87.9 %	82.3 %	83.5 %	72.8 %
②	88.4 %	84.2 %	83.4 %	74.4 %
③	55.3 %	52.8 %	48.9 %	44.9 %
④	83.7 %	80.7 %	63.3 %	64.1 %

4.3. 異なる出品情報文書間における構成要素の抽出精度

本節では，学習データとテストデータに異なる出品情報文書を用いて抽出の精度の検討を行った．用いる素性としては以下の 2 つの組み合わせを設定している．

⑤ 表層文字,文字種,品詞,文節内素性,主辞素性
⑥ 文字種,品詞,分類番号

⑤と⑥の結果をそれぞれ表 4，5 に示す．

表 4 : 表層表現に関係する素性を用いたときの抽出結果

学習データ	テストデータ	属性		属性値	
		適合率	再現率	適合率	再現率
シャツ	キャミ	67.1 %	73.5 %	79.9 %	64.3 %
キャミ	シャツ	88.4 %	57.7 %	84.4 %	58.8 %
キャミ	浴衣	78.9 %	46.5 %	80.6 %	61.1 %
浴衣	キャミ	69.0 %	62.2 %	84.6 %	47.2 %
浴衣	シャツ	87.4 %	62.7 %	86.5 %	44.6 %
シャツ	浴衣	73.7 %	66.5 %	79.4 %	63.9 %

表 5：分類番号を用いたときの抽出結果

学習データ	テストデータ	属性		属性値	
		適合率	再現率	適合率	再現率
シャツ	キャミ	59.6 %	70.8 %	62.1 %	58.0 %
キャミ	シャツ	82.3 %	58.7 %	69.6 %	56.6 %
キャミ	浴衣	67.5 %	51.2 %	60.8 %	60.7 %
浴衣	キャミ	68.4 %	63.2 %	53.5 %	47.0 %
浴衣	シャツ	83.2 %	59.4 %	56.4 %	44.5 %
シャツ	浴衣	57.8 %	66.0 %	48.6 %	60.9 %

表 4，5 の結果より，異なる出品情報文書に適用した場合は同一出品情報文書に適用した場合よりも予想通り精度が下がることが確認された．出現する文字列の違いがあるためであると考えられる．また，分類番号の効果については属性においては表層表現に依存しなくてもある程度の精度が得られることがわかったが，種類が多くなる属性値については，データの組み合わせによって適合率が上がらないという結果になった．やはり，表層表現に依存する部分は大きく今後何らかの検討が必要である．

4.4. 出品情報文書における構成要素の対の同定

本節では，3.4.節で述べた単純な手法を用いて，属性と属性値の対の同定の検討を行った．用いたデータは「アパレル(男性用)-トップス-シャツ-半袖」150 ページである．属性値と属性候補の関係等を示す素性としては，「対象となる属性値の品詞」,「属性候補の品詞」,「属性値と属性候補の間の文字数」,「属性値と属性候補の間の形態素の表層表現」,「属性値と属性候補の位置関係(属性候補が属性値に対して前にあるか後にあるか)」を用いている．表 6 に結果を示す．

表 6：属性-属性値の対の同定の精度

属性候補の探索範囲	対の同定の精度
全範囲	78.0 %
前後 8 配列要素	72.5 %
前後 4 配列要素	83.3 %
前後 2 配列要素	85.4 %
前後 1 配列要素	60.0 %

表 6 の結果より，全体的にある程度の高い精度で対の同定ができることがわかった．属性候補の探索範囲別にみると，前後 2 配列要素から属性候補を取り出しているとき一番精度がよいことから，比較的狭い範囲内に対となる属性と属性値が存在することがわかる．しかし，すべてのペアが近い位置に存在するわけではないのでそれらのペアに関しても対の同定が行えるような手法が必要である．

5. まとめと今後の課題

本稿では，ネットオークションの出品情報文書から商品の特徴的な情報である属性・属性値の自動抽出，対の同定を行うシステムについて述べた．

構成要素の抽出においては，一般的な手法であるチャンキングの手法である程度の精度が得られることがわかった．また，素性として分類番号を用いると表層表現に関係する素性を用いなくてもある程度の精度が確保できることわかった．今後，広いカテゴリ範囲で抽出を行う場合に有効に働くことが期待される．課題としては，出品情報文書を学習処理の前に加工することによっての精度向上，データの量による精度の上昇などが挙げられる．

構成要素の対の同定においては，ネットオークションの出品情報文書では比較的狭い範囲に属性と属性値のペアが存在すると考えられることから，単純な手法でどの程度の精度が得られるかを検討した．精度は一番高い場合で 85.4%とまずまずの精度を示した．精度上昇の手段としては素性の検討や文書の大域的な情報の考慮などが考えられる．

[1] 中野桂吾,平井有三“日本語固有表現抽出における文節内情報の利用” 情報処理学会論文誌,Vol.45, No.3,pp.934-941 (2004)

[2] 森辰則,藤岡篤史,村田一郎 “動向情報編纂のためのテキストからの統計量の自動抽出” 第21回人工知能学会全国大会, 3 H9-4 (2007)

[3] 飯田龍,小林のぞみ,乾健太郎,松本裕治,立石健二,福島俊一“意見抽出を目的とした機械学習による属性-評価値対同定”情報処理学会研究報告,2005-NL-165(2005)

[4] 大野晋,浜西正人 “角川類語新辞典” 角川書店 (1981)

[5] Soon,W.M.,H.T. and Lim,D.C.Y.”A Mzchine Learning Approach to Coreference Resolution of Noun Phrases”Linguistics,Vol.27,No.4,pp.521-544(2001)