ネットオークションの出品情報文書からの 2 段階属性抽出

塚原 裕常† **宮崎 林太郎**† **西村 純**†* **前田 直人**†* **森 辰則**†
小林 寛之‡ **石川 雄介**‡ **田中 裕也**‡ **翁 松齢**‡
†横浜国立大学大学院環境情報学府 〒240-8501 横浜市保土ヶ谷区常盤台 79-7
‡ヤフー株式会社 〒106-6182 東京都港区六本木 6-10-1 六本木ヒルズ森タワー
E-mail: †{ht,rintaro,jun-n,n-maeda11,mori}@forest.eis.ynu.ac.jp
‡{hkobayas,yuishika,yuutanak,shou}@yahoo-corp.jp

１．はじめに

近年の情報抽出技術の利用例の一つとして，ネットオークションにおける出品物の検索の高度化が挙げられる．現在のネットオークションでは，出品物の検索に全文検索が用いられている．このため，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に，図１のように属性情報以外に含まれる文字列に一致する出品も検索されてしまうという問題点がある．しかし，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，それらを高い精度で抽出することが可能となれば，属性検索によって利用者が望む柔軟な検索の実現が期待できる．

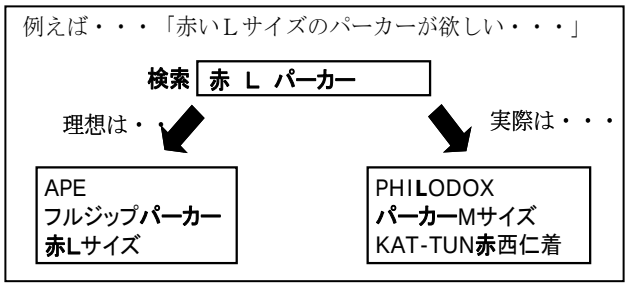

図１：現在のオークション検索の問題点の例

ネットオークションの出品情報文書の例を図 2 に示すが，出品情報文書中には，送料や代金の支払い方法などの記述部分が多く含まれており，出品物の説明をしている記述部分は総テキスト量の半分以下，その中でも出品物の属性情報を含む文となると，総テキスト量の 1/3 にも満たない．そのため，これらの不要な部分をあらかじめ切り分けることにより，属性情報抽出において不要部分に含まれる属性情報に類似した記述の影響が排除できると期待されるまた，属性・属性値の抽出が比較的局所的な文脈情報に基づき行われるのに対して，ある文にそれらが含まれているか否かの判断には文に亘るより大域的な文脈情報が利用できる．よって，両者を組み合わせた手法は，すべてのテキストに対して属性抽出を行う場合よりも精度が向上することが期待される．

そこで，本論文では機械学習手法を用いて，出品情報文書中において，出品物の属性情報が含まれている部分とその他の不要な部分を切り分ける手法を検討する．具体的には，出品情報文書を文単位に分けて二値分類を行うことにより，各文が属性情報を含むかどうかを判定する．

そしてまた，西村ら[1]による既存の属性抽出手法と連結を行い，属性抽出の前処理として本手法が効果的であるかどうかを検討する．

●●●商品説明
【半袖Ｔシャツ黒　サイズＸＬ】新品未着用
男性用ＸＬサイズ対応（胸囲１０４－１１２，身長１７５－１８５）．
１枚目の画像は正面胸部分のロゴ．
２枚目は袖（マーク）と背面（水色部分）です．
お勧めですよ～♪
●●●発送
・定形外郵便（２４０円）
・ポスパケット（４００円，配達追跡が可能です）
●●●お支払い
・ヤフーかんたん決済
・ヤフーネットバンキング

図２：出品情報文書の例

２．関連研究

テキストから必要な情報を抽出する研究の一つに，中野ら[2]が行った日本語固有表現の抽出がある．この研究では文節区切りを行い，文節内の情報を素性としてチャンカーに与える手法を提案している．

また，固有表現抽出の技術を応用して様々なテキストに対して属性抽出が行われている．高橋ら[3]は対象物・属性名・属性値という三つ組の候補を，抽象化したパタンによりドメインを限定せずに抽出し，統計量を使ってフィルタリングする手法を提案している．他にも，新里ら[4]は固有表現抽出手法を用いた，レストランに関する属性情報の自動認識，飯田ら[5]は意見抽出を目的として機械学習に基づく手法を用いた属性と評価値の対の同定方法を提案している．

上記の先行研究に加えて，本研究に最も関連が深い強いものとして，西村ら[1]の研究がある．西村ら[1]は，ネットオークションの出品情報を各種属性により柔軟に検索することを目的として，中野ら[2]の固有表現抽出に基づき，出品情報に多数存在する属性・属性値の情報を機械学習により自動抽出する手法を提案している．

本研究は，西村ら[1]の属性抽出をより正確かつ迅速に行うことを可能にすることを目的としたものである．次節において，西村らの抽出手法を説明し，続く次章において，提案手法について述べる．

２．１ ネットオークションにおける属性抽出[1]

２．１．１ 抽出対象となる属性情報と出品情報文書

西村ら[1]においては，まず，抽出対象となる属性・属性値の定義がなされている．対象とする出品情報文書として，属性情報による検索の需要が多いと思われる，ファッションカテゴリに属するものが用いられていて，表１に示す種類の属性を対象としている．これらの情報はいずれも利用者が検索する際に必要性が高いと考えられるものである．

* 現在，ヤフー株式会社所属

表 1：抽出対象となる属性情報

属性の種類名	出品情報文書中属性値の例
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

２．１．２ 系列ラベリングに基づくチャンキングによる属性情報の抽出

西村ら[1]の手法では，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．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．具体的に与えた素性は，表層文字，文字種，品詞，文節内素性（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する），主辞素性（連続する名詞が存在する場合，その最後の名詞を素性とする），分類番号（角川類語新辞典において各単語に付与されている番号）である．この中でも特に分類番号は，属性抽出の出品カテゴリに対する依存を弱め，カテゴリ横断的な属性抽出を可能にするために用いられた素性である．

また，チャンキングには IOE2 法を用い，チャンキングを行う文字の前後 2 文字ずつ計 5 文字を文脈長としている．

３．提案手法

Kaynak ら[6]は，複数の分類器を直列に多段接続することにより，分類精度を上昇させる手法について提案している．初期の段階においては，単純ではあるが汎用性のある分類器を学習するが，段階が進むにつれて，局所的な事例に特化した分類器を学習する．Kaynak らはこれを Cascading と名づけ，その有効性について考察している．

本論文で提案する手法は，この Cascading の考え方に類似するものである．すなわち，一段階目でより大域的な情報を用いて，大まかな分類を行い，二段目でより局所的な情報を用いて，詳細な抽出を行うものである．具体的には，一段階目として，出品情報文書中の各文についてその文が出品物の属性情報を含むか否かの二値分類を行い，そこで含むと判断された文についてのみ，二段階目の処理である，西村ら[1]による属性抽出を行う．

一方で，Cascading が同質の分類タスクを各段で行っているのに対して，本稿での処理では一段目でのタスクが，実際に行うべき二段目のタスクの範囲を限定するために用いられており，質の異なる分類タスクの組み合わせである点が異なる．

第二段階目については，2.1.2 節で説明したので，本章の残りの部分においては，上記の一段階目について述べる．なお，以下の記述において，「出品情報文書中の当該出品物の属性情報が含まれる文」を「説明記述文」と呼ぶことにする．

３．１ 出品情報文書中の各文に対する素性の付与

本論文では，Support Vector Machine(以下，SVM と記す)を用いた二値分類を行うことにより，出品情報文書から出品物の説明記述となる部分を抽出する．この時，出品情報文書を文単位に分け，文単位で分類を行う．そのため，出品情報文書中の文毎にその特徴量を抽出し，機械学習の際の素性とする．具体的に利用した素性は，以下の３つである．

①：分類対象となる文における形態素の出現頻度
②：分類対象となる文の前後の文の判定結果
③：分類対象となる前後の文の形態素の出現頻度

①は，３つの素性のうち主となるもので，各々の文を形態素解析して得られた形態素原形の出現頻度，いわゆる bag-of-words である．

次に，我々はネットオークションの出品者が出品情報文書を記述する場合，出品物の説明記述を出品物と関係のない文と混ぜ合わせて書くよりは，それぞれをまとめて書く方が一般的ではないかと予測した．これは，出品情報文書には出品物の説明記述の文だけでなく，前述の通り，送料の説明や他の出品物の宣伝などが含まれるためである．

そこで，①のみを素性とした場合の分類結果を初期分類の判定結果として利用し，分類対象としている文の前後数文についての初期分類の判定結果を新たに②の素性として加え，二回目の学習を行う手法についても検討する実験を行った．また，これに加え，同じ理由から，③の素性として前後数文の形態素の出現頻度を採用した加えた場合についても検討する実験を行った．

４．評価実験および考察

出品情報文書中の説明記述文を抽出する実験を行った．形態素解析器には ChaSen を，SVM の実装系としては，TinySVM を用いた．また，この抽出結果を，西村ら[1]の既存の属性抽出システムの入力とする，連結したシステムについて，属性抽出の精度を検証し，提案本手法の有用性について考察する．

評価に際しては，説明記述文の抽出のみの場合，ならびに，連結したシステムの精度評価を行った場合のいずれについても，出品情報文書中の文を要素とした集合を 5 分割して行う交差検定を用いている．また，評価尺度としては，説明記述文の抽出のみの場合は説明記述文が正しく抽出されているかどうかについて，また，連結したシステムの場合は個々の属性・属性値が正しく抽出されているかについて，適合率，再現率，F 値の平均値を求めた．

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-アパレル(男性用)-トップス-シャツ-半袖」（150 ページ・4930 文，属性：総数 1422 個/異なり数 149 個，属性値：総数 1794 個/異なり数 512 個）を用いた．この出品情報文書には，あらかじめ表 1 で示された属性・属性値について XML タグによる注釈付けが行われている．また，出品者固有の記述表現により抽出精度が左右されないために，同一出品者による出品情報は用いないようにした．

４．２ 出品情報文書からの説明記述文抽出

本節では，出品情報文書からの説明記述文の抽出実験について述べる．3 章で述べた①～③の素性を組み合わせた各結果を表 2 に示す．

表 2：説明記述文抽出結果

	適合率[%]	再現率[%]	F 値
①	**90.82**	86.24	**88.47**
①＋②(1 文)	89.41	85.94	87.64
①＋②(5 文)	88.93	86.46	87.68
①＋②(20 文)	89.26	**87.04**	88.14
①＋②(記事全文)	88.04	85.65	86.83
①＋②＋③(1 文)	86.99	85.65	86.32
①＋②＋③(5 文)	87.72	83.67	85.65
①＋②＋③(20 文)	87.53	83.75	85.60

＊括弧内は，素性として用いた前後文の数

表 2 に示す通り，対象となる文における形態素の出現頻度だけを素性として用いた場合でも，適合率，再現率ともにおよそ 9 割程度に至ることがわかった．

属性抽出の前処理ということを考えると，再現率を重視する必要があるが，①＋②，①＋②＋③いずれの場合も①のみの場合と比べて大きく再現率を上昇させることは出来なかった．原因としてはベクトルの次元数が高くなったために，学習がうまくいかなかった，説明記述文とそれ以外の文が我々の予想していた以上に混在していたことなどが考えられる．

４．３ 説明記述文抽出を前処理とした属性抽出

本節では，説明記述文の抽出手法と，西村ら[1]による既存の属性抽出システムとを連結した場合の属性・属性値の抽出実験について述べる．

この時，属性抽出システムの処理対象としては入力には，

学習時データ：コーパス中の注釈情報から判断した，説明記述文(コーパス中の属性もしくは属性値を有する文)の集合．注釈情報つき．

評価時テストデータ：説明記述文抽出手法によって説明記述文であると判断された得られた文．注釈情報なし．

を用いた．結果を表 3 に示す．なお，表中の Baseline は，属性抽出システム単独での抽出結果である．また，「上限値」は，説明記述文の抽出が完全に正しく行われた状況における，属性・属性値の抽出結果であり，説明記述文抽出を前処理に付加することによる精度向上の上限値である．

表 3 によれば，説明記述文の抽出過程を連結することにより属性抽出の精度が，適合率，再現率共に向上しており，素性選択を変更してもいずれの指標もほとんどが Baseline に比べて向上している．Baseline では属性値の再現率が他の指標に比べて 9～14 ポイントも低く，属性値の再現率の低さは Baseline の弱点であったが，説明記述文の抽出過程を連結することによってその弱点が幾分克服できたのではないかと考えられる．

また，Baseline，①，①＋②，①＋②＋③の 4 者を比べると，①はどの指標においても上位の精度を示している．これは説明記述文のみの抽出結果を反映していると考えられる．

表 3：連結したシステムによる属性抽出結果

	属性			属性値		
	適合率[%]	再現率[%]	F 値	適合率[%]	再現率[%]	F 値
Baseline	88.4	84.2	86.2	83.4	74.4	78.6
上限値	93.2	90.3	91.7	91.4	81.9	86.4
①	**89.4**	**87.3**	**88.3**	**87.0**	80.7	**83.7**
①＋②(20 文)	89.2	86.3	87.7	86.2	**81.4**	**83.7**
①＋②＋③(20 文)	**89.4**	83.1	86.1	86.8	79.0	82.8

４．３．１ 説明記述文抽出を連結した場合の属性抽出結果の変化

表 3 のように，説明記述文抽出システムと属性抽出システムの連結により，属性抽出の適合率と再現率はそれぞれ数ポイントずつ向上した．本節では，属性抽出システム単独での抽出結果と連結した場合(4.2 節における素性①を使用)を比較し，抽出結果の変化について考察する．

まず，適合率の向上について考察する．適合率が向上する典型的な状況は，本来属性情報でないにもかかわらず抽出してしまっていた情報が，抽出されなくなることである．今回の連結では，このような例がいくつも見られた．属性抽出システム単独では，「ゆうパック６０サイズ」の「サイズ」や，「汚れが無いことをチェックしておきました」の「チェック」などが属性情報として抽出されていたが，これらは説明記述文抽出の段階で無関係の文を削除することにより抽出されなくなった．これらは，出品物の属性情報の中に同じ表層表現があるため，属性抽出単独では抽出誤りが多くあったと思われる．

次に，再現率の向上について考察する．再現率が向上する典型的な状況は，本来属性情報であるにもかかわらず抽出できていなかったものが抽出されることである．説明記述文抽出を連結した場合では，「素材」や「サイズ」，「デザイン」等に対する属性値を中心に抽出精度が向上している．具体例としては，属性抽出システム単独では抽出できなかった，「豚皮（属性:素材の値）」や「花（属性:デザインの値）」などの属性値が，説明記述文のみに対して属性抽出の学習を行うことで，抽出可能となった．

しかし，連結によって変化したのは必ずしも良い方向にだけではなく，抽出できなくなった属性情報や，誤って抽出してしまうようになった不要な属性情報も存在する．加えて，連結を行ってもなお抽出することができない属性情報も存在することから，説明記述文の抽出過程を前処理として連結する手法は全体的な精度向上をもたらしているものの，局所地的には常に必ずしも精度向上をもたらすものではない．

４．４ 新しい商品の出品情報文書に対する属性抽出

最後に，システムの連結が新しい商品の出品情報文書に対しても有効かどうかを確認するための実験を行った．用いたデータは 4.1 節と同様に 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-アパレル(女性用)-インナーウェア-キャミソール」（150 ページ・4696 文，属性：総数 723 個/異なり数 91 個，属性値：総

数 1245 個/異なり数 512 個),「ファッション-アパレル(女性用)-和服-浴衣」(149 ページ・4682 文，属性：総数 932 個/異なり数 116 個，属性値：総数 1146 個/異なり数 391 個）である．説明記述文抽出の素性には 3.1 節の①を用いた. 抽出対象となる属性情報がすでに定義されているという点からジャンルとしては近い商品の出品情報を扱ったが, 浴衣などは属性の表層表現がシャツと異なる部分も多く，今後扱う出品情報文書が拡大していく場合に，連結の効果が調べられると考えた.

まずは，4.3 節と同様の実験を新たな商品の出品情報文書で行った．結果を表 4 に示す.

表 4：新しい商品の出品情報文書での抽出結果

		属性			属性値		
		適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
キャミソール	Baseline	86.7	74.1	79.9	85.5	69.0	76.4
	連結	**87.1**	**80.9**	**83.9**	**87.6**	**75.9**	**81.4**
女性用浴衣	Baseline	79.9	74.3	77.0	84.2	67.7	75.1
	連結	**82.7**	**81.4**	**82.1**	**85.3**	**78.6**	**81.8**

表を見ると 4.3 節の結果と同様にいずれの場合も精度が向上している．これにより，説明記述文の抽出手法と，属性抽出システムの連結は今後対象とする製品が増えた場合にも有効であることが期待できる.

次に, 学習データとして用いる出品情報文書と，評価データとして用いる出品情報文書を異なる商品の出品情報文書にして実験を行った．これにより，新たな商品が出品された場合に既存データを用いて属性抽出が可能であるかを確認できる.

学習データとして「半袖シャツ」「キャミソール」「女性用浴衣」のそれぞれの出品情報文書を，評価データとしては学習データとは異なるカテゴリの商品の出品情報文書を用いた. これまでの実験は 5 分割の交差検定であったが，ここでは異なるカテゴリの商品の文書を用いているので分割をしていない点が異なる．結果を表 5 に示す.

表 5 を見ると, 説明記述文抽出を前処理として連結することにより，Baseline に比べて再現率が大幅に下がってしまったことがわかる．現在のところ，既存データを用いて新たな商品の出品情報文書から属性抽出を行う際に, 説明記述抽出を前処理として用いるのは有効ではないと言える．これは，商品ごとに現れない表層表現が存在することが原因の一つとして考えられる．そのために，より汎化された素性を追加するなどの検討が必要である.

5．まとめと今後の課題

本論文では, オークションの出品情報文書を対象とした既存の属性抽出について, その抽出精度を向上させる手法を検討した．具体的には，前処理として，出品情報文書の各文に対して出品物の説明記述を含むか否かの二値分類を行うことにより, 送料や関連商品の紹介など出品物と直接関係の無い記述を削除するものである.

説明記述文抽出手法と既存の属性抽出手法とを連結し，高精度の説明記述文抽出が属性抽出の精度向上をもたらすことを確認した．また、新たな商品が登場した際の手法の有効性を確認した. 学習データと評価データが同じカテゴリの商品の場合にはいずれも精度が向上したが, 既存の学習データを用いて新たな商品の属性抽出を行った際には再現率が大きく下がった.

今後の課題としては, 説明記述文におけるさらなる素性の検討，学習データの規模拡大などが挙げられる．また，説明記述文抽出は, 人手によるコーパス作成支援にも用いることができると考えられる.

表 5：学習と評価で異なる商品の出品情報文書を用いた場合の抽出結果

学習データ	商品	手法	属性			属性値		
			適合率 [%]	再現率 [%]	F 値	適合率 [%]	再現率 [%]	F 値
半袖シャツ	キャミ	Baseline	68.2	76.1	71.9	78.6	65.9	71.7
		連結	67.4	59.8	63.3	77.1	52.8	62.7
	浴衣	Baseline	73.7	70.9	72.3	74.0	67.5	70.6
		連結	**79.4**	48.4	60.1	**79.3**	46.6	58.7
キャミソール	シャツ	Baseline	87.5	60.6	71.6	83.0	62.1	71.0
		連結	**90.0**	14.6	25.1	82.7	15.7	26.4
	浴衣	Baseline	80.7	52.1	63.4	78.7	65.3	71.4
		連結	77.7	32.2	45.5	**80.5**	44.3	57.2
女性用浴衣	シャツ	Baseline	86.4	64.5	73.9	87.4	46.5	60.7
		連結	**86.8**	23.6	37.1	31.9	19.9	32.0
	キャミ	Baseline	69.3	70.4	69.9	80.4	51.6	62.9
		連結	66.5	41.2	50.9	**80.7**	30.5	44.2

参考文献

[1] 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，木戸冬子 "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" 言語処理学会第 14 回年次大会発表論文集情報処理学会研究報告, 2007-NL-180, pp392157-395162(20087)

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

[3] 高橋哲朗，乾健太郎，松本裕治 "テキストから属性関係を抽出する" 情報処理学会研究報告，2004-NL-164，pp19-24(2004)

[4] 新里圭司，関根聡，吉永直樹，鳥澤健太郎 “固有表現抽出手法を用いたレストラン属性情報の自動認識" 言語処理学会 第 12 回年次大会 発表論文集(2006)

[5] 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)

[6] Cenk Kaynak and , Ethem Alpaydin “Multistage Cascading of Multiple Classifiers: One Man’s Noise is Another Man’s Data” Proc. of the Seventeenth International Conference on Machine Learning (2000)