社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180 (27)
2007／7／25

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‡ 木戸 冬子‡

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

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

あらまし 本稿では，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，出品情報文書に多数存在する商品の属性，属性値の情報を，機械学習に基づき自動抽出する手法ついて検討している．まず，出品情報の属性検索の対象とすべき属性について考察した．特に，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が少なく，かつ，利用者が検索の対象として欲する属性を抽出対象とした．また，出品情報における属性や属性値の多様性に対応する手法についても考察した．注釈付きコーパスから抽出器を構成する際に，表層表現を直接素性とすると，学習コーパスに特化した学習結果が得られ，特に商品のカテゴリーが異なる未知の出品情報文書からの属性情報抽出の際に精度の低下を招くと考えられるため，表層表現に直接依存しない新たな素性としてシソーラスの分類情報を用い，どのような効果が得られるか検討した．抽出手法としては，固有表現抽出等で用いられる，文字を単位とするチャンキング手法を採用した．

キーワード ネットオークション，属性，情報抽出，チャンキング

Attribute-value extraction from description of exhibits for facetted search in net auction system

Jun NISHIMURA† Rintaro MIYAZAKI† Naoto MAEDA† Tatsunori MORI†

Shorei O‡ Yusuke ISHIKAWA‡ Hiroyuki KOBAYASHI‡ Yuya TANAKA‡ Fuyuko KIDO‡

† Yokohama National University 79-7 Tokiwadai, Hodagaya-ku, Yokohama-shi, 240-8501 Japan

‡ Yahoo Japan Corporation Roppongi Hills Mori Tower, 6-10-1 Roppongi, Minato-ku, Tokyo, 106-6182 Japan

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

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

Abstract In order to achieve flexible facetted search for descriptions of exhibits in net auction system, in this paper, we studied automated extraction of attributes and their values, which appear in those descriptions, based on a machine learning technique. First of all, we examined a set of attributes that should be indexed for the facetted search. Especially, we focused on attributes that can be annotated stably by different annotators, and that are needed for search. We also studied a way to deal with the diversity of attributes and values in descriptions of exhibits. When surface expressions are directly used as one of features, the result of learning may be unwillingly over-fitted to training corpora, and consequently the performance of information extraction will be degraded. Therefore, we introduced the category information of a thesaurus, which does not depend on surface expression directly, and examined the effectiveness of the feature. With regard to the extraction method, we adopted a standard character-based chunking method, which are usually used for named entity extraction.

Keyword net auction, attribute, information extraction, chunking

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」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．図1では色の「赤」ではなく，人名中の1文字である「赤」が抽出されているなどの問題点がある．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，高い精度で抽出する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．本稿では特に属性情報の抽出処理に焦点を当て，考察を行う．

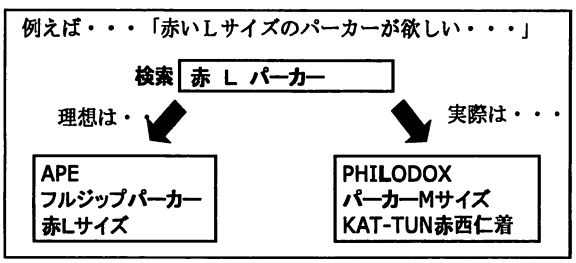

図1：現在のオークション検索の問題点の例

2. 先行研究

まずはじめに，本研究の最終目的でもある属性検索についての先行研究として，廣川ら[3]が教員データに対する多面的検索システムとして，複数の項目からなる教員データを具体的対象として，それぞれの項目を検索の観点として捉える多面的検索方式を提案している．

また，株式会社ジャストシステム[4]の製品 ConceptBase V において，文書に含まれる語句だけでなく，メタデータとして定義されている固有の情報を絞り込み条件として指定することで，XML 文書のタグ情報や属性情報を効果的に活用できるシステムを実現している．

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[5]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[6]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，本稿で取り扱っている属性検索の研究においては，新里ら[7]が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として，レストランに関する情報を既存の固有表現抽出を用いて属性情報の抽出を行なっている．

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た．その際，属性を細かく分類し区別するのではなく，属性と属性値の2種類のみに分類し抽出を行なった．また，一部のカテゴリーに特化しない自動抽出の実現のために角川類語新辞典[8]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抽出精度が得られることを実験により示した．

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，図2に示すように大きく分けて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．

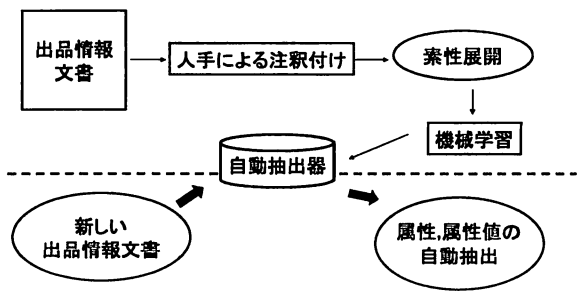

図2：基本的なアプローチ

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ーについて検討を行なっている．

表 1:本研究で定義した属性情報

属性名	出品情報文書中の例
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

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．最終的には，各属性情報に ID 等の情報を付与し組の情報を与え，組の情報を含めて抽出することが理想であるが，現段階では属性には<attr>タグを，属性値には<val>タグを付与するという基本的な問題設定とし、どの程度の抽出精度が得られるか検討する．出品情報文書に対する注釈付けの例を図 3 に示す．

```
::AID::
53916975
::TITLE::
一撃落札！！<val>古着</val>シャツ！！<val>黄</val>
<attr>色</attr> <val>M</val> ボーリング
::DESCRIPTION::
<val>新品</val>！！<attr>サイズ</attr>は、<attr>着丈
</attr>:<val>69</val>、<attr>身幅</attr>:<val>49</val>、
<attr>袖丈</attr>:<val>25</val>、
<attr>肩幅</attr>:<val>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

3.1.2. 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．利用者が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，分解して個別に注釈付けを行う(例:<val>黄</val><attr>色</attr>)．“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分離し，属性，属性値の組が得られたものとして積極的に扱うために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(<attr>サイズ</attr>→<attr>肩幅</attr>，<attr>着丈</attr>，<attr>身幅</attr>など)．

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 節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，シソーラス上の分類番号である．

文節内素性とは先行研究において中野ら[5]が用いた素性で，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．

また，分類番号とは，角川類語新辞典[8]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．例えば「自然」は「天文・暦日・気象・地勢・景観・植物・動物・生理・物質・物象」というふうに，また「行動」は「動作・往来・表情・見聞・陳述・寝食・労役・授受・操作・生産」に，「心情」は「感覚・思考・学習・意向・要求・誘導・闘争・栄辱・愛憎・悲喜」というように分けられている．さらに，これらが 10 個ずつの小項目に分けられている．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．

位置	文字素	文字種	品詞	文節内素性	主辞素性	分類番号	タグ
		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

3.3. 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．

機械学習に基づく抽出手法を利用することを考えると，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．

4. 実験および考察

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に，抽出実験を行なった．また，本研究で素性として新しく用いた角川新類語辞典の分類番号が抽出精度にどのような効果をもたらすかを調べる実験もあわせて行なった．

なお，チャンキングには SVM に基づく汎用チャンカーである YamCha[2]を使用した．文字を単位とし，チャンキングの解析方向は左向き解析で行ない，属性，属性値のチャンクの符号化手法には IOE2 を利用し，文脈長は対象文字の前後 2 文字ずつ計 5 文字とした．

評価に際しては，出品情報を単位とした，5 分割交差検定を行ない，それらの平均の適合率，再現率を求めた．

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カテゴリーのものを用いた．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以下に示す．

- アパレル(男性用)-トップス-シャツ-半袖(150 ページ，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)
- アパレル（女性用）-トップス-タンクトップ,キャミソール(150 ページ，属性：総数 723 個/異なり数 91 個，属性値：総数 1245 個/異なり数 381 個)

上記の出品情報文書に対し，3.1 節で述べた方針に基づき注釈付けした文書を用いた．

4.2. 一般的な素性情報に基づく抽出精度(Case1)

本節では Case1 として，「アパレル(男性用)-トップス-シャツ-半袖」を対象にして，分類番号以外の素性，すなわち，表層文字・文字種・品詞・文節内素性・複合名詞主辞素性を用いて実験を行なった．Case1 は，先行研究でも広く用いられている基本的な素性情報をオークションの出品情報文書で用いた場合どの程度の精度が得られるかを確認するために行なった．

結果を表 2 に示す．

表 2：Case1 における抽出精度

アパレル(男性用)-トップス-シャツ-半袖		
	適合率	再現率
属性	87.8 %	82.3 %
属性値	83.4%	72.9 %

Case1 の条件において，属性に関しては適合率，再現率ともに 80%以上であり，ある程度の精度で抽出できたと考えられる．属性値については，属性に比べて抽出する対象の種類が多くなるために抽出精度は少し下がったと考えられる．これは学習データの量を増やすことで解決することが可能であると考えられるが，それには限界があり，人手による注釈付けの手間もかかるので他の方法を検討する必要がある．

4.3. 表層表現への依存と分類番号素性の効果(Case2)

本節では，機械学習による抽出が表層表現に関係する素性(表層文字，文節内素性，複合名詞主辞素性)を用いた場合，どれくらい影響があるか，また表層表現に依存する素性の代わりに角川類語新辞典の分類番号を用いたときの効果を検討するために以下のように使用する素性を変えた幾つかの条件の下で実験を行なった．表層表現に依存する素性を全く用いないことの利点としては，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，意味的に近い単語が学習用コーパスに存在することによって抽出される可能性があるということである．用いたデータは「アパレル(男性用)-トップス-シャツ-半袖」である．

① 表層文字，文字種，品詞，文節内素性，複合名詞主辞素性(Case1)
② 表層文字，文字種，品詞，文節内素性，複合名詞主辞素性，分類番号
③ 文字種，品詞
④ 文字種，品詞，分類番号

実験結果を表 3 に示す.

表 3 : Case2 における抽出精度

アパレル(男性用)-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	87.8 %	82.3 %	83.4 %	72.9 %
②	88.4 %	84.2 %	83.3 %	74.4 %
③	55.2 %	52.6 %	48.5 %	44.6 %
④	81.0 %	78.1 %	65.4 %	61.5 %

①と②を比較すると，分類番号の効果が僅かだが確認できる．しかし，表層表現に関係する素性への依存が高いため分類番号の効果が少ないことも同時に確認できる．

③と④を比べると明らかだが，表層表現に関係する素性を用いないときには，分類番号は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．

4.4. 学習に用いるオークションデータ量と精度の関係 (Case3)

本節では Case3 として，学習データに用いる出品情報文書の量と抽出精度の関係を調べるための実験を行なった．方法としては，5 分割交差検定法であるが，学習データについては 4 つの学習データすべてから徐々にデータを取り出して，使用する学習データ量を増やし，残りの 1 つの評価用のデータにて精度評価を行っている．データとして「アパレル(男性用)-トップス-シャツ-半袖」を用い，素性には表層文字・文字種・品詞・文節内素性・複合名詞主辞素性・分類番号を使用した．

結果を以下の図 5，図 6 に示す(横軸：オークションページ数[ページ]，縦軸：精度(適合率，再現率)[%]).

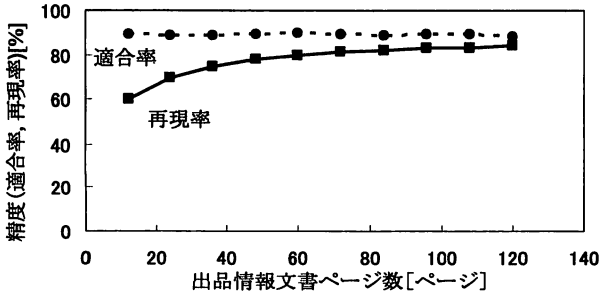

図 5 : 属性におけるデータ量と精度の変化

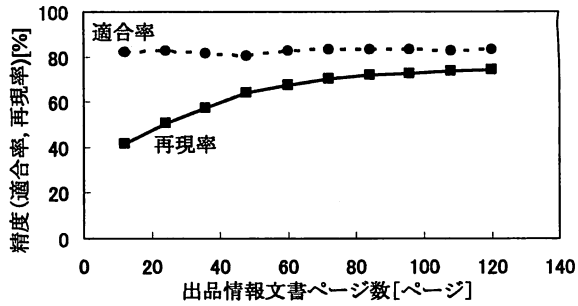

図 6 : 属性値におけるデータ量と精度の関係

図 5，図 6 より，学習データを増加させると適合率の低下を招かずに，再現率を上昇させることが可能であると確認できた．再現率については，上昇が飽和していないためさらにデータを増やすと，向上が見込める可能性がある．

4.5. 異なる出品情報文書間における抽出精度(Case4)

Case4 では，学習データとテストデータに類似してはいるが異なる分野の出品情報文書を用いて自動抽出を行なった場合の精度を検討した．アパレル-トップスの中では，出現する属性情報の類似性の低い「アパレル(男性用)-トップス-シャツ-半袖」と「アパレル（女性用）-トップス-タンクトップ,キャミソール」の 2 つのデータを用いている．ここでは，異なる出品情報文書を用いた自動抽出を行なっているので交差検定法は用いていない．条件を以下に示す．

① 学習データ：アパレル(男性用)-トップス-シャツ-半袖，テストデータ：アパレル（女性用）-トップス-タンクトップ,キャミソール，使用素性：表層文字・文字種・品詞・文節内素性・複合名詞主辞素性

② 学習データ：アパレル(男性用)-トップス-シャツ-半袖，テストデータ：アパレル（女性用）-トップス-タンクトップ,キャミソール，使用素性：文字種・品詞・分類番号

③ 学習データ：アパレル（女性用）-トップス-タンクトップ,キャミソール，テストデータ：アパレル(男性用)-トップス-シャツ-半袖，使用素性：表層文字・文字種・品詞・文節内素性・複合名詞主辞素性

④ 学習データ：アパレル（女性用）-トップス-タンクトップ,キャミソール，テストデータ：アパレル(男性用)-トップス-シャツ-半袖，使用素性：文字種・品詞・分類番号

実験結果を表 4 に示す.

表4：異なる出品情報文書間における自動抽出

	属性		属性値	
	適合率	再現率	適合率	再現率
①	68.1 %	73.9 %	80.0 %	63.9 %
②	60.3 %	71.2 %	62.6 %	57.9 %
③	88.0 %	57.7 %	84.7 %	58.4 %
④	83.1 %	58.7 %	69.2 %	57.1 %

まず, Case1 や Case2 などの同じ出品情報文書で自動抽出を行なった場合と比べるとやはり全体的に精度が低下していることがわかる．これは当然，出現する属性情報の類似性が低くなったからである．

また，①と②，③と④をそれぞれ比較してわかるように，本節で用いた 2 つの出品情報文書間においては表層表現に関係のある素性を用いた抽出精度のほうが高いといえる．ただ，同じ分野の出品情報文書を用いた場合よりも精度の違いの幅は小さくなっていることから，出現する属性情報の類似性がもっと低い場合においてはより有効になってくると考えられるので検討の必要がある．

5. まとめ

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・属性値を自動抽出するシステムについて述べた．また，自動抽出する際に手がかりとなる新たな素性として，角川類語新辞典の分類番号を用いて，出品情報文書における属性や属性値の多様性に対応できるかを検討した．

4.2 節 Case1 では，一般的に使用されている表層文字，文字種，品詞，文節内素性，複合名詞主辞素性という 5 つの素性に基づいた抽出を行ない，ある程度の精度が得られることが確認できた．しかし，この場合は表層表現に関係する素性への依存が大きいと考えられ広い範囲のオークションデータに有効であるかは課題である．

また，分類番号素性の効果について実験を行なった 4.3 節 Case2 では，表層表現が素性として与えられていなくても分類番号を素性として用いると，表層表現，文節内素性，複合名詞主辞素性を素性としたときと比較して精度の低下があまりないことから表層表現に関係する素性に依存しない抽出への適用が考えられる．

4.4 節 Case3 においては，適合率については少ないデータ量である程度の精度が得られるが，再現率については，データ量の増加が精度向上に関係していることを確認した．

4.5 節 Case4 では，異なる出品情報文書間における自動抽出精度の検討を行なった．ここで用いた「アパレル(男性用)-トップス-シャツ-半袖」と「アパレル（女性用）-トップス-タンクトップ,キャミソール」という 2 つの出品情報文書間では，分類番号を素性として用いた場合より，表層表現に関係する素性を用いた場合ほうが精度は高いが，出品情報文書間に出現する属性情報の類似性が低くなるほど分類番号を用いた自動抽出が有効になることも確認できた．

6. 今後の課題と展望

今後，精度を上げるためにいくつかの方法が考えられる．まず，学習データ量の増加による精度の向上である．今回の実験においては，各データ 150 ページ程度の量であるためデータをより増やすと精度が向上するのかについて，検討を行なう必要がある．特に 4.3 節の実験で確認したように再現率において精度向上の見込みがある．

また，出品情報文書においては，出品されている商品とは関係のない事項についての説明文が存在するということがわかっている．その部分を適切に取り除く処理を機械学習の前に行なうことによって必要なテキストだけを残し，精度を上げることを考えている．また，注釈付けの際に不必要なテキスト部分を見る必要がなくなるという作業効率の面からも必要であると考えている．

最後に，本稿ではファッション分野，特にトップスという限定された範囲内で実験を行なっているが，今後出品情報文書の分野を広げていくことが必要である．その際，どのような問題点が出てくるのか，また，本稿で提案した分類番号素性の効果を検討する必要がある．

[1] Asahara, M. and Matsumoto, Y. "Japanese Named Entity Extraction with Redundant Morphological Analysis," Proc. HLT-NAACL 2003 (2003)

[2] Kudo, T. and Matsumoto, Y. "Chunking with support vector machines," Proc. NAACL 2001, pp 1-8 (2001)

[3] 廣川 佐千男，関 隆宏，安元 裕司，山田 泰寛，"教員データに対する多面的検索システム，" 情報処理学会研究報告，2005-DBS-137(Ⅱ)，pp.665-672(2005)

[4] 株式会社ジャストシステム，http://www.justsystem.co.jp/

[5] 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)

[6] 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)

[7] 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)

[8] 大野 晋，浜西 正人 "角川類語新辞典，" 角川書店 (1981)