自然言語処理 132－5
(1999. 7. 22)

広告の自動構造化

井上香織　　横路誠司　　高橋克巳

NTT 情報流通プラットフォーム研究所

住所：東京都武蔵野市緑町 3－9－11
電話番号：0422-59-6646
e-mail：inoue@slab.ntt.co.jp

あらまし　WWW 上の広告情報を効率的に検索するための，自動構造化を目的として，不定型な広告文書内の情報に自動的に属性を付与する方式を研究している．今回，「店・企業名」と「業種」属性について，外部知識から構築した辞書を利用した抽出を行なった．本方式は辞書中単語との完全一致マッチングを基本としているが，辞書のもれを拾うために，識別子マッチングを併用した．店・企業名を表す識別子は，店・企業名データから接頭語，接尾語を切り出して作成した．その結果，辞書マッチングでは見落としていた属性のうち 27% の属性値を拾うことができた．

キーワード　構造化 構造化検索 属性 店・企業名 業種名 識別子

Automatic Structuring of Advertisement Texts

Kaori Inoue　　Seiji Yokoji　　Katsumi Takahasi

NTT Information Sharing Platform Laboratories

Address: 3-9-11,Midori-cho Musashino-shi Tokyo JAPAN
Phone: +81 422 59 6646
e-mail: inoue@slab.ntt.co.jp

Abstract　It's necessary to retrieve data efficiently from large heterogeneous collection of text on the WWW. Our purpose is to establish an automatic structuring method of advertisements texts. We have tried to give attributes to each information in advertisement texts. In this paper, we gave "shop and company name" attributes and " category of business" attributes using the dictionary made with outside knowledge. In order to cover omissions of dictionary matching, our method used distinctive labels to get attribute's values. As a result, label matching could give 27% attributes which could not be given by strict dictionary matching.

key words　structuring　structuring search　attribute　shop/company name　label

広告の自動構造化

井上香織　　横路誠司　　高橋克巳

inoue@slab.ntt.co.jp

NTT 情報流通プラットフォーム研究所

WWW 上の広告情報を効率的に検索するための，自動構造化を目的として，不定型な広告文書内の情報に自動的に属性を付与する方式を研究している．今回，「店・企業名」と「業種」属性について，外部知識から構築した辞書を利用した抽出を行なった．本方式は辞書中単語との完全一致マッチングを基本としているが，辞書のもれを拾うために，識別子マッチングを併用した．店・企業名を表す識別子は，店・企業名データから接頭語，接尾語を切り出して作成した．その結果，辞書マッチングでは見落としていた属性のうち 27% の属性値を拾うことができた．

Automatic Structuring of Advertisement Texts

Kaori Inoue　　Seiji Yokoji　　Katsumi Takahasi

NTT Information Sharing Platform Laboratories

It's necessary to retrieve data efficiently from large heterogeneous collection of text on the WWW. Our purpose is to establish an automatic structuring method of advertisements texts. We have tried to give attributes to each information in advertisement texts. In this paper, we gave "shop and company name" attributes and " category of business" attributes using the dictionary made with outside knowledge. In order to cover omissions of dictionary matching, our method used distinctive labels to get attribute's values. As a result, label matching could give 27% attributes which could not be given by strict dictionary matching.

1.はじめに

1.1 広告検索

WWW上には，多数の企業や店の情報が存在する．一方，情報の受け手は，様々な条件を立てて，アクセスしたい企業や店を選ぼうとする．例えば，火曜日にフカヒレスープが飲める中華料理店を探す．この場合，「業種：中華料理店」「定休日：火曜日以外」「メニュー：フカヒレスープ」に該当するお店を選ぶことになる．現在，サイトによっては，飲食店などある業種の情報を集めて，人手で各店の特徴をデータベース化し，きめこまかい検索サービスを行なっているものもある．しかし，コストがかかりすぎるなどの理由から，全WWW上のデータを扱うことは不可能である．また，全文検索エンジンを使って検索することも考えられるが，例えば先の例の場合，「中華料理」「火曜日」「フカヒレスープ」と入力しても，望んでいるような飲食店の情報が得られる可能性は低い．

ここで，WWW上の情報（ここではテキ

スト情報）に，自動で「業種」や「定休日」「メニュー」といった印をつけて，きめこまかい検索サービスを可能にしようというのが本研究の主旨である．この印をつけることを，一般に「構造化」と呼んでいる．

1.2 構造化とは

次の２つの文はいずれも「カメラ」に関する文である．

１）カメラで写真をとる方法
２）カメラ 大幅値下げ

この２つの文で，単語「カメラ」はそれぞれ異なる使われ方をしている．１）の文では写真をとる「手段としてのカメラ」，２）の文では値下げの対象「商品としてのカメラ」である．このような，文脈毎の単語の使われ方を「属性」と呼ぶ．文書内の情報に属性を付与することで，次のようなことが可能になる．

a)属性を指定した検索が可能になる
b)属性集合から，文書の性質がわかる
c)属性を軸にした情報の統合，再構築が可能になる

ここで，テキスト中の情報（単語など）に属性を与え，属性間の関係を示すことを「構造化(structuring)」と呼ぶ．また，文書を構造化した上で，属性を利用した検索を行うことを「構造化検索」と名づけた．

1.3 構造化のメリット

前節で述べた構造化のメリットについて，本説で詳しく述べる．

a) 属性を指定した検索が可能になる

従来の全文検索では，例えば，「道具」としての「カメラ」と「商品」としての「カメラ」を区別しない．「属性」を考慮せずに，入力されたキーワードにマッチする全ての結果を返すため，無駄な検索結果が多く含まれてしまう．しかし，情報に属性を与えておけば，ユーザは，属性を指定することで，本当に欲しい情報だけを得ることができる．

b) 属性集合から，文書の性質がわかる

文書内にどのような属性の情報が含まれるかを明らかにすることで，その文書の性質を知ることができる．例えば，「住所」と「電話番号」と「店名」が含まれれば，その文書は“広告”という性質を持つ，「日にち（未来）」と「時間」と「場所」が含まれれば，“イベント告知”という性質を持つ，といった推定が可能になる．これは，検索対象の文書集合を分類したり，選択収集するのに役立つ．

c) 属性を軸にした情報の統合，再構築が可能になる

属性を付与することで，その情報は，１文書内の枠を超えて，意味を持つことになる．よって，様々な文書集合から取り出した情報を，属性を軸にして，統合したり，再構築したりすることができる．例えば，A，B２つの文書において，属性「メーカー」と属性「商品名」が同じ「カメラ」があったとする．文書Aの「カメラ」と文書Bの「カメラ」が同じものであることがわかる．このように情報を同定することができれば，文書集合中に属性を軸にしたネットワークを構築することができる．文書という物理的な枠を超えた情報抽出が可能となる．

次節から，対象を広告に絞って，具体的に本研究の説明をする．

2.広告の構造化

2.1 広告からの情報抽出

広告は，その性質上，インターネット上のリソースの中でも，特に情報としての価値が高い．また，ユーザにとっては，属性が付与されていることで，広告を容易に比較することができる．

2.2 必要な属性

広告検索に必要な属性の主なものとして，「住所，緯度経度，電話番号，郵便番号，E-mail Address，時間，URL，店名，業種，商品名，値段」などがあげられる．これらのうち，住所，緯度経度，電話番号，郵便番号，E-mail Address，時間，URL，値段については，筆者らの構造化パーザーを用いて，既に抽出可能であり，属性を与えることができている．
本論文では，「店名」「業種」の属性抽出について述べる．

2.3 属性の抽出方法

属性の付与とは，大きくとらえれば，意味の付与である．よって，統語的意味解析および文脈的意味解析を行った上で意味を与えるのが望ましい．しかし，このような意味解析技術は確立しておらず，処理も複雑になることが想定される．よって，今回

は，厳密な意味解析の手法は用いない．
厳密な意味解析を用いない場合の手法には，

・記号や単語，品詞の出現パターンによるパターンマッチング[1]
「郵便番号」属性 〒XXX－XXXX や，「値段」属性 XXX 円といった表現のバラエティをパターン化しておき，値を抽出する．

・特定分野辞書による単語マッチング
住所辞書に，「新宿」，「新潟」といった単語を登録しておいて，テキスト内の情報とマッチングさせる．

・情報の表現形式が一定の場合，文字列の出現位置を利用する．
例えば，
<CD タイトル> <アーチスト名>
<レーベル名>
<値段>
というように，情報の出現位置が決まっている場合，位置と属性の関係をルール化する．

などが挙げられる．
しかし，これら，パターンや単語，位置からの推定は，文脈を考慮していないため，単独で，高い適合率再現率を保ちながら，全属性を抽出するのは難しい．

本研究で提案する属性付与方式は次の3つの方式を組み合わせたものである．

a) パターンマッチ
b) 辞書マッチ
c) a), b)を行なった上で，他の属性との共起を調べる

既に抽出している属性値のうち，住所を除く属性値は，限られた出現パターンを持つため，a)のパターンマッチで抽出している．例えば，電話番号であれば，XX-XXXX-XXXX，(XX)XXXX-XXXX，といったいくつかの代表的なパターンを用意し，それにマッチした部分を抽出している．
住所に関しては，住所を表す文字列を格納した辞書を用いて，辞書にマッチした文字列部分を抽出している．これは b) の方法にあたる[2].
最後の c)は，抽出した属性の値候補から，より，適合性のあるものを抽出する処理である．例えば，「ＮＴＴインターネットサービス」が，お店の名前らしいとして，本当にそれがお店の名前として使われているかどうかを調べるため，「住所」や「商品名」「値段」といった属性との共起を用いる．共起情報をルール化することで，厳密な意味解析を行わなくても，「NTT インターネットサービス」が属性「店・企業名」と属性「サービス名」を持ちうる（あいまいである）とき，属性を一つに特定することができる．
筆者らの構造化パーザでは，上記の3つの方法を組み合わせることで属性抽出の適合率と再現率を上げている．今回は，以下，「店名」と「業種」の2つの属性について，b)の方法を用いた属性抽出方法の説明と評価について述べる．

3.店名および業種の抽出

3.1 店名の抽出

まず，店・企業名の抽出における課題について述べる．
2節で述べたとおり，属性付与の方法には大きく分けて，パターンマッチと辞書マッチ方式があるが，店・企業名の場合，判別情報は文字の並び方だけなので，記号の位置や文字数を用いるパターンマッチでは抽出できない．よって，辞書を利用した方式を用いる．
辞書を利用するといっても，住所の場合のように，単純に辞書にマッチした文字列を取り出せばいい，というわけではない．主に2つの課題がある．

1． 店名として使用されている単語には，店名特有の単語だけでなく，一般的な単語が多く含まれる．例えば，「猫」，「銀座」「鈴木太郎」など

2． 名前の表現にゆれがある．例えば株式会社○○屋が，○○株式会社と表現されている場合など．

3． 新しい店ができるのは日常茶飯事であり，辞書の更新が追いつかないこと，そうでなくても辞書が全国すべての店名を網羅することは難しいこと，ネット上だけのバーチャルな店が存在することから，店・企業名を固定的な辞書に格納するだけでは対応できない．
つまり，辞書との完全一致マッチング(strict match)のみでは対応できない．

1に関しては，辞書構築の際に，店・企業名特有の単語を優先して登録する方式を提案する．
2に関して，完全一致マッチングでのもれを拾う方法を提案する．

これら2つ以外にも，略称の扱い(日本電信電話株式会社と NTT)や英語表記(マイクロソフトと Microsoft)，といった課題もあるが，今回は扱わない．

3.2 業種の抽出

業種抽出については，店・企業名と業種が対応づけられた外部知識を用いて，店・企業名から業種を導き出す．

3.3 辞書用外部知識 インターネットタウンページ

今回，店の名前とその業種が対応づけられている外部知識として，インターネットタウンページ[3]のデータを用いた．このデータには，日本全国の店・企業名 1100 万件と，業種名約 2000 が対応づけられた形で収録されている．このデータから辞書を構築した．

4.処理概要

4.1 全体像

3.1 節において，1．店・企業名である確率の高い単語だけを辞書に登録する，2．店・企業名の表記のゆれを吸収する，3．辞書に登録されていない店・企業名も拾いたい，という3つの課題について述べた．

これら3つの課題を解決するため，

1. 一般的名詞の排除 そのための基準
2. 店・企業名を判別する「識別子」データの作成

を提案し，実際にこの方法を用いて店・企業名および業種名を抽出した．

以下，提案方法を詳しく説明する．

4.2 一般的名詞の排除と基準

タウンページデータにある 1000 万の店・企業名には，他属性ととられやすい，あいまいなものも含まれる．これらを，店・企業名としての優先度が低いとして削除する．

まず，タウンページのデータから，店・企業名と業種名を対応させた辞書データベースを作成した．（表1参照）

次に，店名と確定できないものの削除を行った．

表2に削除の6つの基準を例とともに示す．（）内は削除数/全店名数．

店・企業名	業種
武蔵野○○郵便局	郵便局
東西南北ＬＥ	電気機械器具製造・卸
いのうえ建機サービス	貸建設用機械器具
株式会社ジュエル	宝石・貴金属加工卸
味処よこじ	居酒屋|飲食店
たかはし左官工業	左官業
△△姫路	自動車販売（外車）

表1 店･企業名辞書

削除基準	店・企業名例	削除数
店名が数値のもの削除	１３	354
固有名詞 (企業名、団体名など以外)	シルビア 一郎	197431
普通名詞削除	電気 猫	221566
人名削除	加藤一郎	141438
１文字以下削除	あ	36963
ひらがな、かたかな、２文字以下	まる マイ	228445

表2 削除基準と数

まず，明らかに属性があいまいであるため，普通名詞，固有名詞，数字を削除した．その後，属性抽出実験の結果，1文字の店・企業名，かな2文字の店・企業名の適合率が極端に低かったため，これも削除した．

以上の処理を行なった上で，826,197 語を削除し，その他を辞書に格納した．

4.3 .店・企業名を判定する「識別子」データの作成

完全一致マッチング抽出できない店・企業名のもれを拾う方法について述べる．

ここでは，単語の出現パターンマッチング手法を応用する．

まず，タウンページのデータから，店名を判別する識別子データを作成する．

識別子とは，「○○工業」「レストラン△△」など，単語の前や後ろに付いて，その単語が，ある属性であることを示すものとする．今回の場合，店・企業名であることを示す接頭語，接尾語を指す．ここで，接頭語と接尾語には2種類がある．識別子と準識別子と呼ぶ．識別子とは，「レストラン」のように，後ろにどんな品詞の単語が接続してもよいもの．準識別子とは，店・企業名を表す識別子であるが，他の属性（商品，サービス名属性）と表現が重なるものである．例えば，「鈴木鉄鋼」の「鉄鋼」は，単独では「商品名」属性ともとれる．この種の識別子は，固有名詞と共起して初めて識別子として働く．他に，「らーめん」「ワイン」「フォト」など．（表3参照）

接頭語	接尾語
アパート・マンション： メゾン	*医院*： 医院 小児科 耳鼻咽喉科 整形外科 科 診療所 病院 皮膚科 クリニック 内科
飼食卸： 海老 マル	*自動車販売(外車)*： 自動車 モータース オート ガレージ
ペットショップ(小鳥)： ズー ペットショップ	*興信・探偵*： 局 調査 本部 本社 リサーチセンター
生活協同組合： 生活協同組合 全国 教職員 生活クラブ生活協同組合 オレンジコープ	

表3 識別子作成前の業種と
接頭語、接尾語の例

各識別子に業種を対応させている．識別子作成手順を次に示す．出現頻度などの統計的手法を用いている．

1.タウンページデータ中の店名を形態素解析（sumomo[4]）する
　居酒屋白糸の滝
　　→居酒屋|白|糸|の|滝
2.接頭語と接尾語を抽出する
　接頭語：居酒屋
　接尾語：滝
3.業種ごとに接頭語，接尾語を分類する
　“居酒屋”　接頭語：居酒屋
　　接尾語：滝
4.各業種につき，出現頻度の高い接尾語，接頭語を採用する
　接頭語：居酒屋　100回出現　→　採用
　接尾語：滝　　3回出現　→　不採用
　今回のパラメータは　5/1000
　　（接語数/ある業種の店名数）
5.多くの業種に同時に現れる接尾語，接頭語は識別能力が低いとして削除する
　株式会社，所，組合，局など
　今回のパラメータは　50/1648
　　（出現業種数/全業種数）
6. 接頭語，接尾語（識別子）を，手作業で普通の識別子と準識別子に分ける．

例えば，表3中の業種“ペットショップ”中の「ズー」は，識別子として適当でないので削除する．“鮮魚卸”中の「海老」は，固有名詞が伴う必要があるので準識別子とし，“生活協同組合”中の「生活協同組合」は，そのままで組織名であると判断できるので，識別子とする．

この結果作成した識別子の数を表4に示す．また，サンプルを表5に示す．

	接頭語	準接頭語	接尾語	準接尾語
異なり識別子数	604	447	969	1470
述べ数 (他業種に出現)	860	653	3130	4384

表4 識別子数

	接頭語	準接頭語
カラオケボックス	カラオケルーム カラオケボックス	カラオケ
ゴルフショップ	ゴルフショップ	ゴルフ

表5 接頭語サンプル

	接尾語	準接尾語
病院・医院 （病院・療養所）	医院 整形外科 クリニック 診療所	
焼肉・ ホルモン料理店	苑 亭 閣	焼肉

表6 接尾語サンプル

5.評価

5.1 実験内容

ロボットで WWW 上から集めた文書中の１００文書をランダムに選び，辞書との完全一致マッチングを行なった．辞書内の単語と，文書内の単語とがマッチしている場合，文書内の単語に対して，「店・企業名」属性を与え，同時に対応する「業種」属性を与える，という方法である．その後，そのもれを拾う形で識別子マッチングを行なった．文書内に識別子があった場合，その前後の品詞の並びから，「店・企業名」属性の値を推定する．識別子には「業種」が対応しているので，同時に「業種」属性を付与する，という方法である．

5.2 評価方法

人手で適合率(抽出した店名候補が本当に店名を表している率)と再現率(文書中の店名が抽出できている率)をはかった．

5.2 結果 （表7，8，9参照）

辞書による全文文字列マッチングだけでは，店・企業名の再現率が3割であったが，識別子マッチングを組み合わせることで5割を超える再現率となった．識別子マッチングの店・企業名の適合率は，全文文字列マッチングに比べ若干下がる．業種の適合率に関しては，かなり差が開いた．これはタウンページで厳密に店名と業種が対応できているためである．店名の適合率は，後に接続する語の品詞を制限したり，ページ中の出現する位置を制限したりすることで上げることも可能であるが，今回は再現率の方を重視した．適合率は，他の属性との共起を用いた処理で上げることができるからである．

全文文字列 マッチング	完全適合	77
	不適合	61
	その他	1
	業種適合	71

表7 全文文字列マッチングの適合数

識別子 マッチング	完全適合	68
	不適合	63
	その他	12
	業種適合	32

表8 識別子マッチングの適合数

	適合率	業種適合率	再現率
全文文字列マッチング	0.56	0.92	0.30
識別子マッチング	0.48	0.47	0.27
全体	0.51	0.71	0.57

表9 全体の適合率と再現率

5.1 不適合，不再現の考察

以下，抽出した店・企業名属性の値のうち，適合していないもの（不適合），再現できなかったもの（不再現），について，その理由を考察する．識別子マッチングに関しては，特有の問題として考察する．

[不適合]

・形態素解析における未知語の問題
タウンページ中の店名のうち，未知語に関しては，店名優先度が推測できないため，削除できない．

例：クリントン(未知語)
ホームページ(未知語)

・店名のようだが，実際は店名ではないもの
例：私のホームページ研究所

－識別子マッチング特有の問題
・表記の問題
例：作者が2行にまたがって書いている
「インターネ ....ットカフェ」
「ットカフェ」になる

・区切りの問題
助詞，記号の入るもの
店・企業名抽出の区切りとして，助詞，記号などを用いているため，店名に助詞が入ると抽出できない．
例：彫刻の森美術館
→ 森美術館となる

[不再現]
・表記の問題
例：作者が2行にまたがって書いている
「インターネ ....ットカフェ」

・ひらがな，カタカナ，漢字，記号の混在「エヌティーティー」(NTT)

・略語表記のもの(英語なども含む)
例：NEC corporation

・店名らしくない，としたもの
例：銀座(バー)

－識別子マッチング特有の問題
・識別子のないもの
例：ピザファクトリー○○
(ピザファクトリーが識別子になっていない)

5.4 適合率の上げ方

2章3節でも少し述べたが，他の属性との共起を用いて適合率を上げる．例えば，「商品名」属性の値を見ることで，業種を推定することができるが，ここで推定された業種と，店・企業名から推定された業種がマッチした場合，その「業種」属性の確からしさは高い，といえる．

6.まとめ

広告構造化を目的として，店・企業名と業種の2つの属性付与方式を提案した．店・企業名の抽出において，辞書マッチング方式を応用して，次の2つの特徴を持つ方式を提案した．
・タウンページデータから，より店名らしいものの抽出して辞書を構築
・タウンページデータから，統計的手法により，識別子データベースを構築
本方式により構造化パーザーを作成し，評価をした．

謝辞 タウンページデータの提供をしていただいたNTT番号情報(株)の各位に感謝いたします．

＜参考文献＞
[1] 桝井等
“新聞記事からの要素属性情報の抽出”
情報処理学会研究報告 98-NL-126-16,
自然言語処理研究会,情報処理学会
[2] 横路等
“特定分野のリソース収集を行うWWWロボットの性能評価”
情報処理学会第57回全国大会
第3分冊,pp163-164
[3]インターネットタウンページ
“http://itp.ne.jp/”
[4] 鷲坂等
“情報検索のための高速日本語形態素解析システム「すもも」”
NTT基礎研究所
第54回情報処理学会全国大会 1997.