# 属性語のWeb文書からの自動発見と人手評価のための基準

徳永　耕亮　　風間　淳一†　鳥澤 健太郎†

本論文では, 広範な概念クラスの属性語を日本語のWeb文書から獲得する手法を提案する．提案する手法は, Web検索を用いて得られた候補の単語を言語的パターン・HTMLタグ・単語の出現の統計量から計算されるスコアで順位付けする簡単な教師無しの獲得手法である．また, 本論文では, 獲得された属性語を人手で評価するための**質問解答可能性**に基づく評価手順を提案する．この評価手順に従い22個の概念クラスに関して提案獲得手法を人手で評価し, 提案手法により属性語を高精度で獲得可能であること, また, スコアに用いた各手がかりが実際に性能に貢献していることを確認した．

**キーワード**:　属性獲得, *Web,* 評価手法

# Automatic Discovery of Attribute Words from Web Documents and Criteria for Human Evaluation

Kosuke Tokunaga , Jun'ichi Kazama† and Kentaro Torisawa†

We propose a method of acquiring attribute words for a wide range of object classes from Japanese Web documents. The method is a simple unsupervised method that ranks candidate words according to the score that uses the statistics of lexico-syntactic patterns, HTML tags, and word occurrences, as clues. To evaluate the attribute words, we also establish an evaluation procedure based on the idea of *question-answerability*. Using the proposed evaluation procedure, we conducted experiments on 22 word classes with four human evaluators. The results revealed that our method can obtain attribute words with a high degree of precision and the clues used in the ranking actually contribute to the performance.

**KeyWords:**　*attribute acquisition, Web, evaluation method*

## 1　はじめに

我々の物の理解の仕方に関する知識は多くの自然言語処理タスクにおいて重要である．物をどのような観点から理解するかということを述べる**属性**の知識はその一つである．例えば，「車」の属性は「重量」,「エンジン」,「ハンドル」,「操作感」,「製造会社」などである．言い換えれば，属性とは，我々があるものについて知りたいときにそれに対する値（本論文の言い方では,「答え」）が知りたくなるような項目である．従って，属性知識の応用としては，情報の要約 (Yoshida, Torisawa, and Tsujii 2003, 2004)，質問応答 (Fleischman, Hovy, and Echihabi 2003;

† 北陸先端科学技術大学院大学情報科学研究科, School of Information Science, Japan Advanced Institute of Science and Technology

高橋, 乾, 松本 2004) などが考えられる．また，最近では機械学習や単語クラスタリングの際の素性として有用であることも示されている (Almuhareb and Poesio 2004)．このような属性知識は，WordNet (Fellbaum 1998) のように人手で作成することも可能であるが，作成コストとカバレッジが問題となる．本研究では，これらの問題を解決するため，与えられた概念クラスの**属性語**[1]をWebから自動獲得する手法を提案する．

属性語の自動獲得を目指した研究はそれほど多くはない．既存研究には，質問応答を念頭において⟨対象, 属性, 値⟩という事実の集合を獲得しようとするもの (Fleischman et al. 2003; 高橋他 2004) や, 情報要約の際に副産物的に属性的な単語を生成するもの (Yoshida et al. 2003, 2004) などがあるが，概念クラスの属性語を明示的に獲得し，その精度を詳しく評価したものはなかった．我々は，属性知識の段階での問題の性質を明らかにし，属性語をあらかじめ高精度で獲得しておくことが，最終的には質問応答などのために値まで獲得する場合などでも大きく役に立つという考えから，属性語の獲得に焦点をしぼる．

属性語は語彙知識の一つと言える．これまで語彙知識の自動獲得としては，上位下位関係の獲得 (Hearst 1992; Shinzato and Torisawa 2004)，全体部分関係の獲得 (Berland and Charniak 1999)，言い換え関係の獲得 (Barzilay and McKeown 2001) などが試みられてきた．上位下位関係や全体部分関係など名詞間の関係の獲得に関しては，目的の関係を特異的に示す言語的あるいは書式的なパターン，その他の統計的な手がかりを相補的に用いて獲得するアプローチがある程度の成功をおさめている (Hearst 1992; Berland and Charniak 1999; Shinzato and Torisawa 2004)．以下で概要を述べるが，本研究で提案する獲得手法もこの範疇に入る．

本研究で提案する獲得手法では，クラス$C$（例えば,「車」）の属性語を獲得するために，まず，$C$を含む文書をWebから検索エンジンを用いて発見し，収集する[2]．収集された文書から属性語の候補を抽出し，それらを言語的パターン・HTMLタグ・単語の出現に関する統計値を利用したスコアに従って順位付けし，スコアの高い候補を属性語として出力する．このスコアは，属性語に関する我々の観察が反映されるように設計されている．前述したように，言語的パターンは他の語彙知識獲得手法でも用いられてきた (Fleischman et al. 2003; Almuhareb and Poesio 2004; Hearst 1992; Berland and Charniak 1999; 高橋他 2004)．特に，本研究で用いる言語的パターンは,「$C$の$A$」という助詞「の」を介したパターンである（ただし，$A$は属性語候補）．このパターンは，直感的に有用と考えられ，関連研究である (高橋他 2004) でも同様のパターンが用いられている．また，属性知識の特殊な場合である全体部分関係を英語を対象として獲得した (Berland and Charniak 1999) でも「$A$ of $C$」という類似したパターンが用いられている．

この獲得手法の新規性は，広範なクラスに対して属性語を獲得することを目的としてWebを情報源として用いること，その際，クラスと関連の高い文書に注目するためWeb検索を用い

1 本研究では，属性が実際に言語で表現される時の文字列を属性語と呼ぶ．テキストからの自動獲得では，実際に獲得できるのは属性語であり，複数の属性語が同じ属性を表すことがあり得るが，これらの認識は本研究の対象外とする．
2 本論文では，混乱が無いと思われる場合には，クラスとクラスを表す語（クラス語）の両方を$C$と表記する．

ること，それにともない，HTMLタグといったWeb特有の手がかりを利用できることにある．ただし，手法はできるだけ簡素になるようにした．標準的な言語パターンを用い，頻度やdf・idfなどの単純な積をスコアとして用いる．また，正解データの作成はコストがかかることから，(Fleischman et al. 2003)のような教師付き学習を用いるアプローチではなく，教師無しで獲得することを目指した．実験では，この提案手法で各クラスに対して上位20個の属性語を出力した時に，約73%の適合率で厳密な属性語が獲得でき，約85%の適合率で緩い属性語が獲得できることを示す[3]．

属性語獲得の研究では，属性語の定義，言い換えれば，獲得された属性語に対する評価基準が確立されていないことも問題になる．本研究では，質問解答可能性という考えに基づいた言語テストによる評価手順を示すことで，この問題の解決を目指す．属性語を定義するには，例えば「もし$A$が，$o$をクラス$C$に属するインスタンスとした場合に$v = A(o)$のように関数的に働き，$v$が$o$をクラス$C$の他のインスタンスから区別するのに重要であるならば，$A$は$C$の属性語である」のように分析的に定義することも可能であるが, このような分析的な定義は人手の評価で直接用いるには複雑で難しく，評価結果の信頼性も低くなると予測される．そこで，本研究では，いくつかの簡単な言語テストを用いた評価方法を提案する．言語テストは，評価者の直感を利用したYES-NOテストであり，評価者の負担が軽減され評価結果の信頼性も向上すると考えられる．提案する評価方法は「属性とは答えが知りたくなるような項目である」という我々の元々の直感を反映したもので,「その値を問うような質問文を生成でき，それに対して答えが存在するならば属性語である」という考え（質問解答可能性）に基づく．本研究ではこの考えに基づいた評価手順を設計する．

属性語の判定のための言語テストはこれまでにも提案されている． 例えば，Woodsは「the $A$ of $o$ is $v$」という表現が可能かどうかで判定できることを述べている(Woods 1975)．しかし，この言語テストを自動獲得された属性語の評価に実際に適用した研究はこれまで行われていない． また，本文で詳しく述べる通り，この基準だけでは，特に日本語に置き換えたときに，重要でない語が属性語と判定されてしまうなどの誤判定が発生する可能性がある．本研究で提案する判定方法は，質問解答可能性の考え方に基づいた言語テストによって，より重要な属性語に焦点をあてるとともに，いくつかの補足的な言語テストを組み合わせることで，より正確な判定を目指したものである．

最後に，いくつかの文献が指摘する通り，属性には「重さ」などの性質,「エンジン」などの部分,「操作感」などのtelic的属性,「製造会社」などのagent的属性など多くのサブタイプがある(Guarino 1992; Pustejovsky 1995)．しかし，これらの区別が無いとしても，属性は前述した応用で有用であり，また，区別のための評価基準は複雑で安定した評価が困難になるということから, 本研究ではこれらの区別は無視することにした．

3 厳密な属性語・緩い属性語の違いについては本文で詳細を述べる．

本論文の構成は以下の通りである．節2で，属性語獲得のための提案手法の詳細を述べる．次に，節3で属性語の評価基準とそれに基づく評価手順を示す．節4で，提案手法を提案評価手順で評価した実験の結果を示し, 節5でいくつかの考察と今後の課題を述べる．

# 2　獲得手法

この節では，属性語の自動獲得手法の詳細を述べる．

## 2.1　属性語の性質に関する観察

はじめに，獲得手法の基になった属性語の性質に関する我々の観察を示す．具体的には，属性語には以下に挙げる三つの性質があることが分かった．

**性質 (1)**　属性語は，助詞「の」を含む「$C$ の $A$」という言語的パターンでクラス語と共起する傾向がある．

**性質 (2)**　属性語は，Web 文書中で HTML タグを用いて強調表示されたり，リストや表の要素として出現する傾向がある．

**性質 (3)**　属性語は，クラス語を含む文書に出現しやすく，他の文書にはあまり出現しない傾向がある．

以下では，これらの性質を利用した獲得手法を提案する．

## 2.2　属性語候補の獲得

提案手法では，まずはじめに，属性語の候補となる語を以下のように Web から収集する．クラス $C$ の属性語を獲得する場合，クラス語 $C$ を含む文書を Web 検索エンジンを用いて求め，ダウンロードする．本研究ではこの文書集合を局所文書集合 (local document set) と呼び，$LD(C)$ と表記する．このような収集の方法は前節で述べた属性語の性質 (3) を反映していると考えられる．次に，この $LD(C)$ 中の全ての名詞[4]を取り出し，これを属性語の候補とする．複合語が属性語になる可能性もあるが，簡単のため，本研究では一語からなる属性語のみを扱うことにした．

## 2.3　属性語候補の順位付け

前節の方法で得られた属性語候補は，節2.1で述べた属性語の性質のうち性質 (3) を考慮しているとはいっても，属性語でない語も多く含んでいる．そこで，属性語候補を他の性質も反映したスコアによって順位付けし，上位の語のみを属性語として出力するようにする．本研究で提案するスコア関数はいくつかのサブスコアを掛け合わせた以下の形をしている．

4 形態素解析器 JUMAN(Kurohashi and Nagao 1999) によって普通名詞・サ変名詞・地名・未定義語（のうちカタカナかアルファベット）と判定された語である．

**表 1**　スコア$n(C, A)$のための言語的パターン

| | | | | |
|---|---|---|---|---|
| $C$の$A$は | $C$の$A$を | $C$の$A$から | $C$の$A$で | $C$の$A$へ |
| $C$の$A$が | $C$の$A$に | $C$の$A$まで | $C$の$A$より | $C$の$A$、 |

$$V(C, A) = n(C, A) \cdot f(C, A) \cdot t(C, A) \cdot dfidf(C, A). \quad (1)$$

$A$は属性語候補であり，$C$はクラスである．$n(C, A)$と$f(C, A)$は言語的パターンに関するサブスコアで性質(1)を反映している．$t(C, A)$はHTMLタグに関するサブスコアで性質(2)を反映している．$dfidf(C, A)$は単語の出現に関する統計値によるサブスコアで，性質(3)を反映したものである．これらのサブスコアを掛け合わせることで，正しい属性語に高いスコアが与えられることを期待している．サブスコアの組み合わせ方には，他にいくつも選択肢が考えられるが，ここでは最も単純な方法の一つを選択した．

以下では，これらのサブスコアの詳細を述べる．

## 2.4　サブスコアの詳細

まず，$n(C, A)$は性質(1)を反映したスコアである．性質(1)で述べたように，属性語の獲得には助詞「の」を介して$C$が$A$に係る言語的パターン「$C$の$A$」が大きな手かがりになると期待される．従って，$n(C, A)$としては$C$と$A$が「の」を介して係った回数などが考えられる．本研究では，$n(C, A)$として$C$と$A$が局所的文書集合$LD(C)$中で表1に挙げたパターンのいずれかで共起した回数を用いる．「$C$の$A$」の係り受けの回数は「$C$の$A$P」(Pは助詞あるいは句読点)というパターンの出現回数である程度近似できると考えられるからである[5]．

$f(C, A)$は33年分の係り受け解析済みの新聞記事中[6]で$C$と$A$が「$C$の$A$」という係り受けで共起した回数である．$n(C, A)$に加えて$f(C, A)$を用いる理由は，マッチさせる文書の量を増やして信頼性の高いスコアを得るためである．本研究を進める過程で，我々が用いたものも含め商用の検索エンジンではクエリにマッチする文書のごく一部のURL（最大で一千文書程度）しかユーザに提示しないという制限があり，現実にはクラス語を含む文書を大量に収集できないという問題があることが分かった．解決策としては，表1で挙げたパターンでのフレーズ検索のヒット件数を用いることも考えられたが，実験で述べるように属性語候補の数は２万程度になり，商用検索エンジンに大きな負担をかけることになる．そこで，本研究では我々が既に持っていた大量の係り受け解析済み新聞記事を用いることにした．将来的には，検索結果の取得に制

5 助詞あるいは句読点の存在によって，例えば「$C$の$A$N」(Nは名詞)のように$A$が複合名詞の一部になっている場合を間違ってカウントしてしまうことを防ぐことができる．

6 読売新聞　1987年–2001年，毎日新聞　1991年–1999年，日経新聞1983年–1990年　(全体で3.01 GB) を日本語係り受け解析器(Kanayama, Torisawa, Mitsuishi, and Tsujii 2000)で解析したもの．

<B>タイ風・カレー</B><BR>材料<BR>鶏肉400g，なす2個，バイマックルー 2枚，ナンプラー大さじ1.5<BR>赤唐辛子1.5本，砂糖小さじ1，ココナッツミルク，バジル<P>スパイス<BR>コリアンダー，クミン<P>作り方<BR><OL><LI>材料をペースト状にして，カレーペーストを作る</LI><LI>カレーペーストを熱した鍋に加えて香りを …

**図 1**　HTML文書の例

限のないWebリポジトリを構築し，大量のWeb文書から共起回数を求める予定である．

$t(C,A)$ は $A$ が $LD(C)$ 中にHTMLタグで囲まれて出現した回数，より正確には，「<*tag1*>$A$<*tag2*>」という形式で $A$ が出現した回数である． ただし，回数をカウントする際には，HTMLタグ間の文字数(つまり $A$ の長さ)は最大でも20と制限する．長い文字列は単語ではなく文になっていることが多く，ほとんど属性にはなり得ないからである．また，ここでの<*tag1*>と<*tag2*>は開始タグ(<A>など)・閉じタグ(</A>など)のどちらでもよいことにする．例えば，図1のHTML文書では,「タイ風・カレー」「材料」「スパイス」「コリアンダー，クミン」「作り方」などの語がカウントされる．このスコアは属性語の性質(2)を反映したもので，Web文書中で強調表示される語，改行などによりフォーマットされる語，列挙や表の要素になる語などに高い値を与えることを目的としている．

最後に，$dfidf(C,A)$ は性質(3)を反映したスコアである．このスコアでは $A$ を含む文書に多く出現し，しかも，特徴的な語（ストップ語のような普遍的な語でない語）に高い値を与えることが目標である．本研究では，上位語と下位語の関連度を測るために新里ら(Shinzato and Torisawa 2004)が用いたスコアを参考にして，以下の式で計算されるスコア関数を用いることにした．

$$dfidf(C,A) = \quad df(A, LD(C)) \cdot idf(A), \quad idf(A) = \log \frac{|G|}{df(A,G)}.$$

ここで，$df(A,x)$ は文書集合 $x$ 中で $A$ を含む文書の数を表している．$G$ は大域的文書集合(global document set)と呼ばれるランダムに収集された大量のWeb文書であり，Web全体を近似した文書集合である．これを用いて $idf$ を計算することによりWebにおける特徴的な語を知ることができる．

## 3　属性語のための評価基準

自動獲得された属性語の良さを評価するには，何らかの定まった評価手順が必要になる．本研究では，我々が質問回答可能性と呼ぶ基準に基づいた評価手順を提案する．

質問回答可能性とは，ある語に関して，その値を問う質問文を生成でき，答えが存在する，ということである．我々は，質問回答可能性が成り立つならばその語は属性語であるという仮

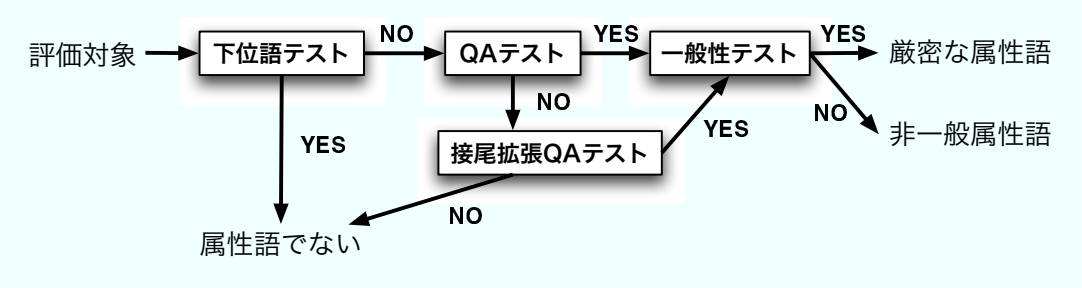


**図 2**　属性語の評価手順

**下に挙げる質問文の中に，文法的に正しく，常識的に自然で，**
**答えが仮想的にでも想像できるものはありますか？**

1. この$C$ の$A$は何?　4. この$C$ の$A$はどこ?　7. この$C$ の$A$はどう?
2. この$C$ の$A$は誰?　5. この$C$ の$A$はどれ?
3. この$C$ の$A$はいつ?　6. この$C$ の$A$はいくつ?

**図 3**　QAテスト（$C$はクラス語，$A$は判定しようとしている属性語）

定をした．これは，我々の属性の利用方法（QAや要約）を考えると，直感的に妥当な仮定だと思われる．例えば,「車」について考えると,「この車の製造会社はどこか？」といった質問文が可能であり，誰かが「A社」と答えることができる(質問回答可能性が成り立つ)ので,「製造会社」は属性語と考えても良いということになる．ある人が「車」について知りたいとき「製造会社」が何であるかは重要な情報であるので，これは妥当であろう．

提案する評価手順では，評価者は最大で４つのYES-NO質問に答えることで判定を行う．これら４つの質問とは，評価者に提示される順に，下位語テスト(節3.4)，QAテスト(節3.1)，接尾拡張QAテスト(節3.2)，一般性テスト(節3.3)，である．QAテスト・接尾拡張QAテストの２つが，質問回答可能性を直接用いたテストである． 下位語テスト・一般性テストは，QAテストを補強するためのものである．評価手順の全体の流れは，図2に示すようになる．

以下の節では，各テストの詳細を説明する．

## 3.1　QAテスト

実際の順番とは異なるが，まず，本研究の提案の中心的な判定手順である質問回答可能性テスト(QAテスト)から説明する．我々は，前述した質問回答可能性基準に従って，図3に示すQAテストを設計した．

このQAテストについて，いくつか注意点がある．第一に，属性語に対する値は実際にはクラスのインスタンスに対して定義されるため，このQAテストでは，$C$を「この」で限定することで（ある）インスタンスを指すようにして質問が不自然になるのを防いでいる．また，質問文中で$C$の後にスペースを入れることで評価者が適切な係り受け（「この$C$」が$A$に係る）を想像できるようにした．

第二に，$A$に対する適切な質問の種類をあらかじめ（自動で）決めることは難しいので，テストでは図3に挙げたように考えられる全ての種類の質問を生成し，そのいずれかが許容できるかを判定するようにした．

第三に，質問は文法的に正しいだけでなく「常識的に自然」でなければならない．本研究の実験では，この「自然」さを「その質問が通常の会話の第一発話としてあり得るか」で判断するようにした．我々の考えでは，属性語はインスタンスを述べる際に重要なものでなければならない．そこで，この「自然」さを満たす属性がそのような重要な属性語であるという仮定を置いた．例えば，おそらく全ての「会社」は「机」を所有しているが，我々の考えでは「机」は「会社」の属性語ではない．「この会社の机は何ですか？」といった質問は文法的に正しいけれども，「会社」についての通常の会話の第一発話としては不自然であるから，この「自然」さのチェックによって「机」が属性語になるのを防ぐことができる[7]．重要な点は，Woodsの言語テスト (Woods 1975)(「the $A$ of $o$ is $v$」が可能か) だけでは「the desk of (= used in) Com-X is Desk-Y」のように言えてしまうので，「机」を棄却することができないことである．また，質問は質問者が重要であると考えることについてするのであるから，質問文を判定に用いることでより重要さを重視することができる．

最後に，質問への答えは必ずしも言語で表現できなくてもよい．例えば，「地図」「姿」「設計図」などに対する値は言語では表現できず，他の手段で表現されるが，これらも重要な属性語であることは明らかである．Webページには言語以外の表現（画像，音声など）を含めることができるため，それらを値とする属性への言及も増えると考えられる．そのため，Webから属性語を獲得する場合，そのような属性語も獲得される可能性が大きい．従って，質問への答えが言語に限らないことをあらかじめ評価者に明確にしておいた．

## 3.2　接尾拡張QAテスト

獲得された属性語のいくつかは，正しいと思われるにもかかわらず前節のQAテストで棄却されてしまうことがある．大きな理由の一つは，獲得された属性語が，実際に意味している属性の標準的な属性語とは異なる文字列として獲得される場合があることである．これは，日本語が省略的であること，我々の獲得手法が実際にコーパスに現れた表層形を処理して獲得すること，また，1単語の属性語しか獲得しないことなどに起因している．例えば，下の例文中で

7 もちろん，「机」は事務機器の販売員にとっては「会社」の重要な属性語かもしれないが，本研究ではあくまでも普通の人々にとって通常の状況で重要な属性語を獲得することを目標とする．

- 「$A$(の)$S$」 ($S$=数，方法，名，者，時間，時刻，時期，場所，金額，程度，具合)
- 「$A$の$Y$さ」 ($Y$さ=形容詞・形容動詞の名詞化「高さ」「重さ」など)

**図 4**　接尾拡張QAテストで許される拡張（$A$は元の属性語）

「生徒」は属性「生徒数」の意味で用いられている．

(a) この学校の生徒は５００人です．

このような文を手がかりにすると，提案手法では「生徒」が「学校」の属性語として獲得される可能性が高い．一部が省略されている属性語でも，実際に(a)のような文で使用されることから応用の面で有用であると考えられるので，正しいと判定されるのが我々の立場では好ましい．ところが，省略があると前節のQAテストで棄却されてしまうことがある．例えば，上の「生徒」を判定する場合，これが「生徒数」を意味しているときには, QAテストの質問の中では「この学校の生徒はいくつ？」が一番適当であるが，これは，人数に「いくつ」は使えないため文法的ではない．そのため,「生徒」は棄却されることになる[8]．

日本語では，省略された部分のほとんどは属性語の後に適切な接尾辞を付加するか,「の＋名詞化形容詞・形容動詞」を付加することで復元することができる．例えば，上の例では接尾辞「数」を付ければ良い．そこで，最初のQAテストで棄却された場合には，適切な付加を行って属性語を拡張した上で，それを評価すべき語としてQAテストを再度行うようにした．上の例では,「生徒数」をQAテストで再び評価する．実際のテストで許される拡張は，図4に示した通りである．可能な接尾辞については，図4に示した適用範囲が広いと思われるものに限定し，これでカバーできないものは，名詞化した形容詞・形容動詞を適切に付加してもらうことで対処した[9]．このテストを，接尾拡張QAテストと呼ぶことにする．

## 3.3　一般性テスト

我々の当初の目的は与えられたクラスに対する属性語を獲得する，つまり，クラスの全てのインスタンスに共通の属性語を獲得することであった．しかし， 評価者によっては，全てのインスタンスに共通の属性語ではないが興味深い属性語を正しいものとして判定することが予備実験において分かった．例えば，クラス「映画」に対する「字幕」や「車」に対する「後席」などである．全ての映画に字幕がある訳ではないので，厳密には「字幕」は映画の属性語ではない（例えば，日本では字幕はほとんど外国映画に対して付与される）し，全ての車に後席がある訳ではないので厳密には「後席」は「車」の属性語ではない．しかし，厳密に属性語では

8 「生徒」を個々の構成員を表す属性として判定するとしても,「この学校の生徒は誰？」は文法的であっても「常識的に自然」の要請を満たしていないとして棄却されてしまう可能性がある．

9 接尾拡張QAテストは，初めのQAテストでの疑問文が網羅的であれば必要なかったかもしれない． しかし，我々は，簡単で限定的なテストで大部分をカバーし，カバーしきれないものをより複雑なテストで再確認するほうが評価の負担が減り評価の安定性が向上すると考えた．

| $A$と$C$の間に、以下に挙げる関係のどれか一つでもなりたっていますか？ |
| --- |
| 「$A$は$C$の一種である」「$A$は$C$のひとつである」「$A$は$C$の一人である」 |

**図 5**　下位語テスト

なくてもそれを持つようなインスタンスの割合・重要性が高い場合には，正しいと評価される傾向があることが推測される．このような属性語はその割合・重要性から実用的に有用であると考えられる．このような全てのインスタンスに共通する属性語とそうでない属性語の性質を調べることができるように，QAテストで受理された属性語に関しては，それがそのクラスの「ほとんど全て」のインスタンスに共通するかを最後に判定するようにした．このテストを一般性テストと呼ぶ．一般性テストで受理された属性語を「厳密な属性語」と呼び，QAテストで受理されて一般性テストで棄却された属性語を「非一般属性語」と呼ぶ．また,「厳密な属性語」と「非一般属性語」を合わせて「緩い属性語」と呼ぶことにする．実験では，厳密な属性語としての獲得精度，緩い属性語としての獲得精度を調査比較する．

### 3.4　下位語テスト

最後に，下位語テストについて説明する．我々の提案手法では，誤って獲得された属性語の中にクラス$C$の下位語やインスタンスと考えられる語が多く含まれることが分かった．もし，評価対象の$A$が$C$の下位語やインスタンスなら，それは$C$の属性語にはなり得ないが,「$C$の$A$」という表現が自然になってしまうためQAテストで混乱を引き起こしやすい．例えば,「アニメ」のインスタンスとして「ドラゴンX」があるとすると,「アニメのドラゴンX」という表現は自然であり,「ドラゴンX」は明らかに「アニメ」の属性語でないのにQAテストで誤って受理されてしまう可能性がある．そこで，QAテストの前に図5で示される下位語テストにより$A$が$C$の下位語やインスタンスであるかを判定し，下位語やインスタンスでない場合にだけQAテスト以降に進むようにした．

逆に，$A$が$C$の上位語である場合には，上位語であっても必ずしも$A$が属性語でないとは言えないのでそのようなテストは行わない[10]．

## 4　実験

この節では，提案獲得手法を前節で述べた評価手順で評価した実験について述べる．

10 例えば,「アニメ」に対する「映像」のように,「映像」は「アニメ」の上位語であるが,「このアニメの映像はどう？」-「きれい」などの質問回答ができるので属性語である．上位語であって属性語でない場合にはQAテストでほとんど棄却できる．

**表 2**　評価で用いた22個のクラス

| |
|---|
| 都市，博物館，祝日，警察，施設，大学，新聞，ごみ，神社，鳥，病院，植物，川，小学校，曲，図書館，支店，サイト，町，センサー，研修，自動車 |

## 4.1　実験設定

まず，評価のために32個のクラスを用意した．Webに現れるようなクラスで評価を行うため，この32個のクラスは，新里らの上位下位関係獲得手法(Shinzato and Torisawa 2004)によってWebから獲得された共通の上位語をもつ単語クラス1,589個の中から選んだものであり，このときの上位語をクラス語として用いる．単語クラスに含まれる下位語は，クラス語の意味が曖昧な場合に意味を特定するための情報として評価者が参照できるようにした．また，我々の目的は上位下位関係獲得の評価ではないので，上の獲得手法でうまく獲得されている単語クラスを選んだ．さらに，この32個のクラスからランダムに22個のクラスを選び(表2)，評価の対象とした[11]．

提案獲得手法で用いられる局所文書集合$LD(C)$の収集にはWeb検索エンジンであるgoo (http://www.goo.ne.jp)を用いた[12]．$LD(C)$の大きさはクラス平均で857文書(URL)であった．また，この$LD(C)$から得られた属性語候補はクラス平均で約2万語であった．サブスコア$dfidf(C,A)$の計算に必要な大域文書集合$G$としてはWebからランダムに収集した$10^6$文書を用いた[13]．

各クラスについて提案手法および後で述べる比較手法による上位50個の属性を出力し，評価対象とした．効率的に評価するため，上記の全ての手法による属性語を一つの集合にまとめ(重複があれば取り除く)，評価の公平性を保つためランダムに並べ替えた．このように重複を除くと，評価すべき属性語は全てのクラスで合わせると3,678個だった．これらの属性語を，提案評価手順を実装したGUIツールを用いて4人の評価者がそれぞれ4日間かけて評価した．この評価結果から，それぞれの手法の上位50個の出力に対する評価が生成できる．

この実験に関して，評価者間の一致度を示すkappa値(Landis and Koch 1977)は，厳密な属性語の評価としては0.533，緩い属性語の評価としては0.593となり，両者とも「中程度」の一致を示した．

## 4.2　提案手法の精度

図6に提案獲得手法による緩い属性語としての精度，図7に厳密な属性語としての精度を示す．

それぞれの図において左のグラフは各評価者（Evaluator1-4）による適合率，右は評価者に

11 評価にかかる時間・コストの制約のためにこのような選択を行った．
12 $C$が検索エンジンの形態素解析により分割されてしまうのを防ぐため，フレーズ検索（完全一致）で検索している．
13 これは，論文(Shinzato and Torisawa 2004)で用いられた文書集合と同じものである．

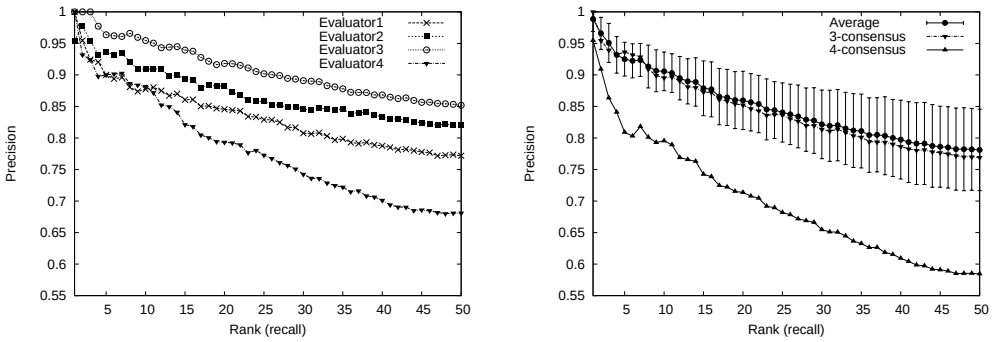


**図 6**　緩い属性語としての精度

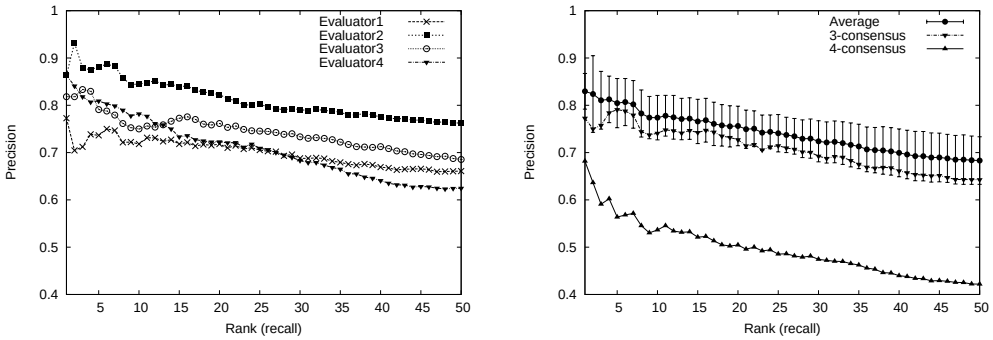


**図 7**　厳密な属性としての精度

関する平均・3 人一致（3-consensus）・4 人一致（4-consensus）の適合率である[14]．グラフの $X$ 軸は上位何個まで集計するかを，$Y$ 軸はそのときの適合率を表している．大まかに言って，グラフの $X$ 軸は再現率に対応する[15]．上位 $n$ 個における評価者 $k$ による適合率 $P_k$ は，$\mathcal{C}$ を評価に用いたクラスの集合とすると，

$$\frac{1}{n|\mathcal{C}|}\sum_{C\in\mathcal{C}}(C\text{の上位}n\text{個の出力の中で}k\text{が正しいと判定した属性語の数})$$

で計算される．評価者に関する平均の適合率は，$\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}P_k$ で計算される（$\mathcal{K}$ は評価者の集合）．また，$M$ 人一致の適合率は，

$$\frac{1}{n|\mathcal{C}|}\sum_{C\in\mathcal{C}}(C\text{の上位}n\text{個の出力の中で}M\text{人以上が正しいと判定した属性語の数})$$

14 平均に関しては縦棒で ± 標準偏差を示す．

15 出力されるべき全ての属性を知ることはできないので再現率を正確に計算することはできない．これらのグラフは，正確な再現率を $X$ 軸とした場合より，適合率が高い手法に多少不利なグラフになっている．

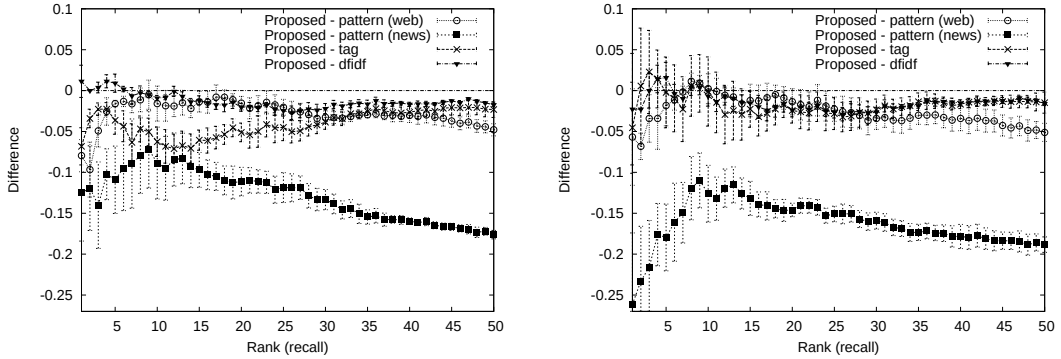


**図 8**　各サブスコアの効果．左: 緩い属性語の場合 右: 厳密な属性語の場合．
$n(C, A)$ の効果は「Proposed - pattern (web)」，$f(C, A)$ の効果は「Proposed - pattern (news)」，
$t(C, A)$ の効果は「Proposed - tag」，$dfidf(C, A)$ の効果は「Proposed - dfidf」で示す．

で計算される．

グラフをみると，適合率自体は評価者に大きく依存するが，提案手法の順位付けと適合率の間の正の相関は共通して存在することが分かる．これは提案手法の妥当性をある程度示していると言える．また，緩い属性語としての評価と厳密な属性語としての評価を比べると，厳密な属性語の方が獲得が難しいことが分かる．加えて，厳密な属性語としての評価（つまり，一般性テスト）は評価者によって大きく傾向が違うことが分かる．緩い属性語の評価の場合にはプロットはほとんど交差していない． これは評価者間で許容度の差はあっても評価の傾向は変わらないことを示している．一方，厳密な属性語の場合には，プロットが交差しており， 評価者によって評価の傾向が異なることを示している．緩い属性語の場合に一番許容的であった評価者3（図中，Evaluator3）が厳密な属性語の場合にはそうでもない点などは興味深い．これらのことから，一般性テストは他のQAテストなどに比べて一致を得るのが難しいテストになっていることが推測され，先に示したkappa値の違いもそれを示唆している．

提案手法による獲得精度はおおむね期待の持てるものである．どの評価尺度を採用するのが妥当かは難しい問題であるが, 例えば(Berland and Charniak 1999)で用いられた多数決基準（本実験の場合， 3人一致）を用いるとすれば，提案手法は上位20個の属性語を出力した場合，緩い属性語は0.852，厳しい属性語は0.727の適合率で獲得できることになる．表3に，実際に獲得された属性語の上位20個をいくつかのクラスに対して示す．これらをみると，実際に興味深い属性語が獲得できていることが分かる．

## 4.3　各サブスコアの効果

次に，提案手法のスコア(式1)における各サブスコアの効果を調べるため，式1から各サブ

**表 3**　提案手法による上位20個の属性語.
括弧中前の数字は緩い属性語として判定した評価者の数,
後の数字は厳密な属性語として判定した評価者の数である.

| クラス | 属性語 |
|---|---|
| 鳥 | 写真[4/4] 名前[4/2] 種類[4/4] イラスト[3/3] 特徴[4/4] 病気[4/2] 生活[4/4] 話題[3/2] 関係[0/0] イメージ[4/4] 巣[4/4] 鳴き声[4/4] 姿[4/4] 情報[4/4] 世界[0/0] 声[4/4] 動物[0/0] ページ[3/2] 生態[4/4] 羽[4/4] |
| 病院 | ホームページ[4/1] 施設[3/3] 情報[4/4] 紹介[4/4] 窓口[4/4] 認定[3/3] 名称[4/2] 医師[4/4] 精神科[4/2] 評判[4/4] 対応[4/4] 電話[2/2] 診療[4/4] 治療[4/4] 医療[3/3] 機能[3/3] 院長[4/4] 評価[4/4] 診察[4/4] ページ[2/2] 管理[4/3] 一部[1/1] |
| 植物 | 名前[4/2] 種類[4/4] 写真[4/4] 種子[4/4] 栽培[4/3] 観察[4/3] 特徴[4/4] 説明[4/4] 画像[4/4] 調査[4/3] データ[4/4] 進化[3/3] 解説[4/4] リスト[2/2] 葉[4/3] 保存[2/2] デザイン[1/1] 生育[4/4] |
| 川 | 水位[4/4] 上流[4/4] 名前[4/2] 環境[4/4] 水質[4/4] 歴史[4/4] 源流[4/4] 写真[4/4] 水[4/4] 水面[4/4] 場所[4/4] 流れ[4/4] 水辺[4/4] 水源[4/4] 四季[3/3] 特徴[4/4] 中[1/1] ほとり[4/4] 自然[4/4] せせらぎ[4/4] |
| 小学校 | 活動[4/4] 取り組み[4/3] 運動会[4/4] 子ども[4/4] ホームページ[4/0] 校長[4/4] 教室[4/4] 校歌[4/4] 児童[4/4] 校舎[4/4] 行事[4/4] 学習[3/3] 給食[4/3] ページ[2/2] 体育館[4/4] 学級[3/3] メール[0/0] 学年[1/1] 始業式[4/4] 音楽[2/2] |
| 曲 | 歌詞[4/1] タイトル[4/2] 演奏[4/4] リスト[0/0] イメージ[4/4] 作詞[4/1] 楽譜[4/4] 名前[4/2] 内容[3/3] ジャンル[4/4] 情報[4/4] ポイント[4/4] 世界[1/1] メロディー[4/4] 最後[3/2] 題名[4/2] 中[0/0] 作曲[4/4] テーマ[4/4] データ[4/2] |
| 図書館 | 資料[4/4] ホームページ[4/2] ページ[3/1] 歴史[4/4] 設置[4/4] システム[4/4] 蔵書[4/4] コピー[2/2] 本[4/4] 場所[4/4] 利用[4/4] サービス[4/4] データベース[4/3] 図書[4/4] 新聞[4/4] 休館[4/4] 目録[3/3] 展示[4/2] 施設[2/2] 情報[4/4] |
| 支店 | 所在地[4/4] パソコン[2/1] 紹介[4/4] 歴史[3/3] 営業[4/3] 電話[2/2] ホームページ[4/1] 住所[4/4] 窓口[4/3] 駐車場[4/3] |
| サイト | 情報[4/4] 掲示板[4/2] 内容[4/4] 運営[4/4] リンク[3/2] 登録[3/2] 紹介[4/3] 写真[2/1] 中[1/1] コンテンツ[4/4] |
| 町 | 人口[4/4] 歴史[4/4] ホームページ[4/0] 観光[4/4] 情報[3/3] 財政[4/4] 施設[4/4] 文化財[4/2] 環境[4/4] 温泉[3/1] 話題[3/2] 四季[3/3] イベント[4/3] 図書館[4/3] 文化[4/4] 風景[4/4] シンボル[4/3] 産業[4/3] 農業[4/2] 議会[3/3] |
| センサー | 情報[4/4] 感度[4/3] 種類[4/3] 位置[4/4] 取り付け[4/4] 開発[4/4] 精度[4/4] サイズ[4/4] 仕様[4/4] 温度[2/1] データ[4/4] セット[4/4] 設置[4/4] 機能[4/4] 技術[4/4] 特長[4/4] ページ[3/3] 高さ[3/2] 採用[3/3] 応用[4/4] |
| 研修 | 内容[4/4] 目的[4/4] 実施[4/4] テーマ[4/3] プログラム[4/4] 講師[4/4] 予定[4/4] 名称[4/2] メニュー[4/4] 報告[4/4] 対象[4/4] 成果[4/4] 充実[2/2] 場[3/3] あり方[2/2] 詳細[4/4] 機会[1/1] 定員[4/4] 受講[4/4] ほか[0/0] |

スコアを除いたスコアを用いたときに精度がどの程度変化するかをみた．まず，評価者平均精度の変化でみると，全てのサブスコアに関して効果がある（各サブスコアを除く事で精度が低下する）ことが観察された．さらに，各評価者ごとにみると，評価者によらず似た傾向の精度の変化があることが分かった．これは，前の実験で示したように評価自体は評価者によってかなり異なる事を考えると興味深い．この点をより詳しく分析するため，各サブスコアを除いた場合に関して，評価者ごとに精度の変化を計算し，変化の評価者に関する平均・標準偏差など

を求めた．図8はそのようにして求めた平均・標準偏差をプロットしたものである．このグラフから，全般的に，ほぼ順位にかかわらず，全ての評価者で効果があることが分かる．左のグラフで示された緩い属性語の場合を詳しくみると，$f(C,A)$ と $t(C,A)$ の効果が特に大きい．また，$n(C,A)$ は $f(C,A)$ と同様に効果はあるが絶対値は小さいことが分かる．これは，前述したように利用できる文書の量の差によるものと考えられる．$dfidf(C,A)$ は，低ランクまででみると正の効果を示しているが，高ランク域（1位–5位）ではわずかではあるが負の効果になっている．右のグラフの厳密な属性語の場合にも，全般としては正の効果があることが分かる．しかし，効果の程度は全般に小さくなり分散も大きくなっている．特に，$t(C,A)$ の効果は緩い属性語の場合に比べて大幅に小さくなっている．一方で，$f(C,A)$ の効果は逆に大きくなっている．

### 4.4　局所文書集合収集のためのキーワードの比較

提案手法では，順位付けで用いられる局所文書集合を収集する際の検索エンジンに対するキーワードとしてクラス語を使用した．しかし，もし上位下位関係知識が利用できる場合には，クラスに属する下位語をキーワードとして局所集合を収集する事も可能であり，その場合に属性語の獲得精度がどのように影響されるかは興味深い問題である．そこで，この実験では提案手法のようにクラス語を用いる場合と下位語を用いる場合を比較した．実験で用いたクラスには，上位下位関係獲得手法 (Shinzato and Torisawa 2004) により下位語が対応付けられているので，これを検索キーワードとして用いた．ここでは，収集された文書の質を比較するため，収集された文書から提案手法で収集された文書数とほぼ同数の文書をランダムで取り出して用いた．このようにして得られた局所文書集合を用いてスコア中の $n(C,A)$ を計算する際には，$C$ の代わりに下位語 $H$ を用いた場合に得られる頻度を全ての下位語について和をとって $n(C,A)$ の値とした (つまり，$n(C,A) = \sum_H n(H,A)$)．他のサブスコアの値については提案手法と同じ値を用いた．図9に，前節の実験と同じ方法で緩い属性語と厳密な属性語について提案手法と比較手法の精度の差（差の平均 $\pm$ 標準偏差）を示す．負の差は比較手法の精度が提案手法に比べて悪いことを表す．従って，この結果から，少なくともこの設定では局所文書集合の収集にはクラス語のほうが適していることが分かる．

## 5　今後の課題

本研究では，提案手法によってある程度の高精度で属性語を自動獲得できることを示したが，本格的な応用のためには，獲得精度のさらなる向上が必要である．また，属性語の性質についても更なる考察が必要である．以下に挙げる点が，今後の課題として考えられる．

**質問回答可能性に基づいた手がかり**　提案手法が現在用いている順位付けのスコア（式1）は，節3で述べた評価基準の背後にある質問回答可能性などの考えを直接は反映していない．

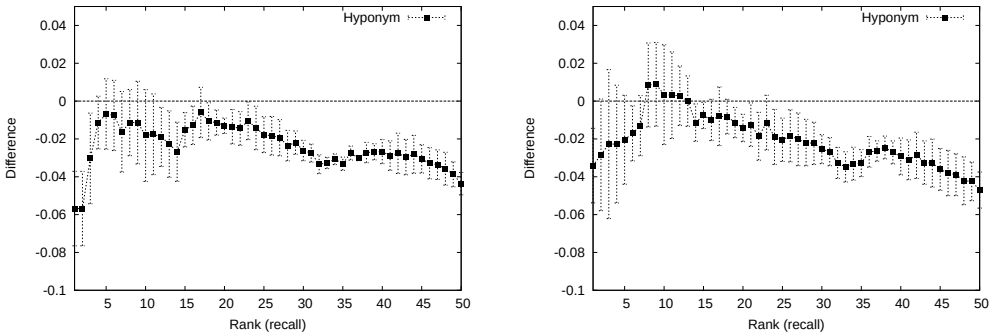


**図 9**　クラス語と下位語の比較．左: 緩い属性語の場合．右: 厳密な属性の場合．

$n(C, A)$ あるいは $f(C, A)$ などで,「$C$ の $A$」という言語的パターンの頻度としてわずかに反映されているだけである．属性語を提案評価手順で評価できるものと仮定するならば，獲得手法でも質問回答可能性などを手がかりとして直接利用することでより高い精度を達成できると考えられる．質問回答可能性を直接反映させるためには，例えば，Web 上の FAQ ページから得られる統計値を使うことなどが考えられる．また，上位下位関係のデータベースが利用できるならば，評価手順中の下位語テストを反映したようなスコアを設計する事も可能である．実験では，厳密な属性語の獲得の精度が緩い属性語の獲得の精度に比べて低いことが分かったが，順位付けのスコアに属性語の一般性を直接捉えるようなサブスコアがないことを考えると，ある程度予測できることである[16]．この場合にも，上位下位関係のデータベースを用いれば，例えば，下位語の何割にその属性語が当てはまるかなどの統計値を用いて属性語の一般性を反映したようなスコアも設計できると考えられる．

**Web の最大限の利用**　現在の提案手法では，前で述べた通り検索エンジンの制限から Web の文書を完全には利用できていない．利用できる Web 文書が増えればサブスコア $n(C, A)$ の信頼性が上がり精度向上に役立つと考えられる．現在我々は，独自に収集した Web 文書に対して制限のない検索エンジンを構築することで，これを実現させることを計画している．大量の Web 文書があると下位語の過疎性も軽減されるので，節 4.4で述べた上位語（クラス語）と下位語の有用性についてもより詳しく分析することが可能になる．

**実応用に向けた再現率の調査**　獲得された属性語，また，本研究で提案した評価基準の妥当性は，究極的には実応用でどれだけ有用かによって判断される．実応用では，必要な属性語のどの程度が獲得できるかという属性語の再現率も重要になってくる．本研究ではこの点については分析していないので，例えばあるクラスについて考え得る属性語を人手

16 検索キーワードや言語パターン中でクラス語を用いる事で一般性が間接的に捉えられると考えられるが，クラス語だけでは曖昧性などの問題も起きていると考えられる．

で列挙し，そのうちどれくらいが実応用で必要になるかなどの分析を行いたいと考えている．また，本研究では上位下位関係獲得手法でうまく獲得できたクラスを評価に用いた．そのため，比較的易しい(頻繁に現れる)クラスについてのみ評価している可能性は否定できない．そこで，より難しい（稀な）クラスについての評価も必要になる．

**属性の型の獲得**　ある属性に対してどの質問が可能か，どのような接尾拡張が可能か，言い換えると，属性の値にどのような語が可能か（属性の型）という知識は，最終的に〈対象,属性,値〉の組まで獲得したい場合や，属性の種類（性質なのか全体部分かなど）を決定したい場合に重要になると考えられる．本研究では，獲得の際にはこの点を無視し，評価の際には評価者の判断に任せていた．今後は，このような知識も提案手法のような単語や言語的パターンの統計値を用いた方法などを用いて獲得したいと考えている．

## 6　結論

本研究では，Webから言語的パターン・HTMLタグ・単語の統計量を手かがりとして属性語を獲得する手法を提案した．また，獲得された属性語を評価するための質問回答可能性に基づいた評価手順を提案した．この評価手順を用いて提案獲得手法を評価し，属性語を高精度で獲得できること，また，用いた各手がかりが精度に貢献していることを確認した．

**謝辞** 実験で使用したデータに関してアドバイスをいたただいた新里圭司氏に深く感謝いたします．また、実験で評価者として参加していただいた北陸先端科学技術大学院大学の学生の皆様に感謝いたします．

## 参考文献

Almuhareb, A. and Poesio, M. (2004). “Attribute-based and value-based clustering: An evaluation.” In *Proc. of EMNLP 2004*, pp. 158–165.

Barzilay, R. and McKeown, K. R. (2001). “Extracting paraphrases from a parallel corpus.” In *Proc. of EACL 2001*, pp. 50–57.

Berland, M. and Charniak, E. (1999). “Finding parts in very large corpora.” In *Proc. of ACL '99*.

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. The MIT Press.

Fleischman, M., Hovy, E., and Echihabi, A. (2003). “Offline strategies for online question answering: Answering questions before they are asked.” In *Proc. of ACL 2003*, pp. 1–7.

Guarino, N. (1992). “Concepts, attributes and arbitrary relations: Some linguistic and ontological criteria for structuring knowledge base.” *Data and Knowledge Engineering*, **8**, 249–261.

Hearst, M. A. (1992). “Automatic acquisition of hyponyms from large text corpora.” In *Proc. of COLING ’92*, pp. 539–545.

Kanayama, H., Torisawa, K., Mitsuishi, Y., and Tsujii, J. (2000). “A hybrid Japanese parser with hand-crafted grammar and statistics.” In *Proc. of COLING 2000*, pp. 411–417.

Kurohashi, S. and Nagao, M. (1999). “Japanese morphological analysis system JUMAN version 3.61 manual.”.

Landis, J. R. and Koch, G. G. (1977). “The measurement of observer agreement for categorial data.” *Biometrics*, **33**, 159–174.

Pustejovsky, J. (1995). *The Generative Lexicon*. The MIT Press.

Shinzato, K. and Torisawa, K. (2004). “Acquiring hyponymy relations from Web documents.” In *Proc. of HLT-NAACL04*, pp. 73–80.

Woods, W. A. (1975). *Representation and understanding: Studies in cognitive science*, chap. What’s in a link: Foundations for semantic networks. Academic Press.

Yoshida, M., Torisawa, K., and Tsujii, J. (2003). *Web Document Analysis*, chap. Chapter 10 (Extracting attributes and their values from Web pages). World Scientific.

Yoshida, M., Torisawa, K., and Tsujii, J. (2004). “Integrating tables on the World Wide Web.” *Transactions of the Japanese Society for Artificial Intelligence*, **19** (6), 548–560.

高橋哲郎, 乾健太郎, 松本裕治 (2004). “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　自然言語処理　2004-NL-164, pp. 19–24.

## 略歴

**徳永 耕亮:**　2003年日本大学工学部機械工学科卒業．2005年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．修士（情報科学）．同年，(株) 日立製作所入社．

**風間 淳一:**　1999年東京大学理学部情報科学科卒業．2004年東京大学大学院情報理工学系研究科コンピュータ科学専攻博士課程修了．博士（情報理工学）．同年，北陸先端科学技術大学院大学情報科学研究科助手．

**鳥澤 健太郎:**　1992年東京大学理学部情報科学研究科卒業．1995年同大学大学院理学系研究科情報科学専攻博士課程退学，同年より同専攻助手．1998年より2001年まで科学技術振興事業団さきがけ研究21研究員兼任．2001年より北陸先端科学技術大学院大学情報科学研究科助教授．計算言語学の研究に従事．博士（理学）．

(1995 年 5 月 6 日 受付)
(1995 年 7 月 8 日 再受付)
(1995 年 9 月 10 日 採録)