HTML 文書からの単語意味クラスの単純な自動獲得手法

新里圭司[†] 鳥澤健太郎[††]

本稿では意味的に類似した自然言語表現の集合である単語意味クラスを，HTML 文書から高い精度で高速に獲得する手法を提案する．Shinzato らによれば，HTML 文書中の表・箇条書きなどの構造には単語意味クラスと見なせる表現の集合が含まれると報告されている．しかしながら，すべての表や箇条書きが意味的に類似した表現の集合を含んでいるわけではない．そこで本研究では，既存の検索エンジンより得られるヒット件数と，それを基に計算される相互情報量を素性とする Support Vector Machine を用いて，表や箇条書きに含まれる表現間の意味的な一貫性を求める．このとき，本手法では n 個の表現を含む表・箇条書きに対しては，$2n$ 回検索エンジンに問い合わせるだけで意味的一貫性の計算を行う．提案手法により獲得された単語意味クラスを 4 人の被験者により評価した．その結果，入力として与えた表・箇条書きのうち，意味的一貫性の高い上位 10%を単語意味クラスとして獲得した場合，その 8 割が 4 人中 3 人の被験者により単語意味クラスとして判断された．

A Simple WWW-based Method for Semantic Word Class Acquisition

Keiji Shinzato[†] and Kentaro Torisawa[††]

This paper describes a simple method to obtain semantic word classes from HTML documents on the web. Shinzato and Torisawa previously showed that itemizations in HTML documents can contain semantically coherent word classes. However, not all the itemizations are semantically coherent. Our goal is to provide a simple method to extract only semantically coherent itemizations from HTML documents. Our method can perform this task by obtaining hit counts from full text search engines $2n$ times for an itemization consisting of n items. The obtained hit counts are used for calculating mutual information values between items in an itemization, and the hit counts and mutual information values are given to a Support Vector Machine as features. The itemizations are ranked by using the Support Vector Machine, and highly ranked itemizations are produced as semantically coherent word classes. In our experiments using four human subjects, when the top 10% of given itemizations were produced, at least, three of the four human subjects regarded about 80% of the produced itemizations as semantically coherent classes.

1. はじめに

本稿では，World Wide Web（WWW）上に大量にある HTML 文書から意味的に類似した自然言語表現の集合（たとえば，{ 薩摩宝山，西海の薫，喜六，七窪 }[☆]）を高速に自動獲得する手法について述べる．本研究では，**要素どうしが意味的に類似している単語または複合語の集合**を**単語意味クラス**と呼ぶ．

従来より，新聞記事などのコーパスを対象に単語意味クラスの自動獲得に関する研究は数多く行われてきた[1),3)~8),10)]．しかし，そのほとんどは名詞と名詞，または名詞と動詞といった単語間の共起関係を利用するものであり，高い精度で単語意味クラスを獲得するためには，構文解析などの深い解析が必要となる．このような従来手法を，WWW 上に膨大な量存在する HTML 文書に対して適用し，大量の単語意味クラスを獲得しようとする試みは，計算時間の問題があるため現実的ではない．たとえば，河原ら[13)] によれば，WWW 上より収集した HTML 文書 4 億件に対して形態素解析および構文解析を行うだけで，およそ 10 カ月かかると試算されている[☆☆]．また，たとえ高速に形態素解析や構文解析が行えたとしても，従来手法ではきめ細かい単語意味クラスの獲得が難しいという問

† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†† 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（JAIST）

☆ これらは芋焼酎の銘柄である．

☆☆ 河原らは大規模なクラスタを利用し，短時間で形態素・構文解析を行っているが，そのようなクラスタは誰でも利用できるわけではないため，計算時間の問題は残る．

題もある．たとえば「薩摩宝山」と「闇魔」を，従来手法により「芋焼酎」と「麦焼酎」のような異なるクラスに分類することは，両者とも共起する名詞ないし動詞が酷似しているため難しいと考えられる．しかし，このようなきめ細かい単語意味クラスは，実用的なアプリケーションを考えた場合に有用である．

一方で，構文解析を主用せずに単語意味クラスを獲得する手法として，Shinzato らが提案した上位下位関係獲得手法[9] がある．この手法は，HTML 文書中の表や箇条書きに含まれる表現の集合を下位語の集合と見なし，各下位語に共通する上位語を求め上位下位関係を獲得する手法である．共通の上位語を持つ下位語の集合は意味的に類似しており，単語意味クラスと見なせることから，この手法は一種の単語意味クラス獲得手法と見なせる．しかしながら，Shinzato らの手法においても，HTML 文書のダウンロードや，ダウンロードした HTML 文書の部分的な係り受け解析といった比較的重い処理を必要とするため，高速に単語意味クラスを獲得できるとはいえない．

以上より，WWW 上に大量にある HTML 文書から単語意味クラスを高い精度で獲得することを考えた場合，構文解析や HTML 文書のダウンロードを必要とする従来手法では，スケーラビリティや，獲得される単語意味クラスの「粒度」という点で，十分とはいい難い．

従来手法のかかえる上記の問題点を解決するため，本手法では，(1) HTML 文書の構造，(2) 既存の検索エンジンより得られるヒット件数の 2 点を利用して，単語意味クラスの獲得を行う．本手法では，まず，同一の表や箇条書きに含まれる表現の集合を単語意味クラスの候補として抽出する．次に，既存の検索エンジンより得られるヒット件数と，それを基に計算される相互情報量[1] を素性とする Support Vector Machine[11] を利用して，抽出された単語意味クラス候補の**単語意味クラスらしさ**を表すスコアを求める．そして最後に，スコアの高い候補を単語意味クラスとして獲得する．以上の処理を経ることで，きめ細かい単語意味クラスを高速に獲得することが可能になる．これは，たとえば酒類を WWW 上で販売しているホームページでは，「薩摩宝山」と「闇魔」が表や箇条書きなどで「芋焼酎」と「麦焼酎」という観点から分類されやすいこと，および「薩摩宝山」と「闇魔」のヒット件数は既存の検索エンジンを利用して手軽に得られることを考え合わせると容易に想像できる．

また，単語意味クラスの統一的な定義を与えることは難しいが，本研究では既存のシソーラスを利用して単語意味クラスの定義を行った．そして，この定義に従って，提案手法が獲得した単語意味クラスを 4 人の被験者により評価した．その結果，表・箇条書きに含まれる表現の集合のうち，単語意味クラスらしさを表すスコアの上位 10%を出力とした場合，少なくとも 4 人中 3 人の被験者によって，その 8 割が単語意味クラスであると判定された．

本稿の構成は以下のとおりである．まず 2 章で関連研究について述べた後，3 章で提案手法について述べる．続いて 4 章で単語意味クラスの定義および，実験結果について報告し，5 章で本研究のまとめを行う．

2. 関 連 研 究

HTML 文書中の表や箇条書きから単語意味クラスを獲得する研究としては，先述した Shinzato らの上位下位関係獲得手法（以降，Hyponymy Relation Acquisition Method を略して HRAM と呼ぶ）[9] が考えられる．HRAM では，すべての表や箇条書きが共通の上位語を持つ表現を含んでいるわけではないため，獲得された上位語と表や箇条書きから抽出された表現（下位語）の係り受け関係と，ヒューリスティックルールを用い，妥当な上位語が獲得されやすい表・箇条書きから優先的に出力するという一種のフィルタリングを行っている．このフィルタリングには，検索エンジンを利用して表や箇条書き中の要素を含んでいる HTML 文書を 1 要素あたり 100 件収集し，そこから各要素の係り受け関係を得るという比較的重い処理を要する．提案手法は，HRAM のように表や箇条書き中の各表現に共通する上位語を求めることはできない．しかしながら，検索エンジンより得られるヒット件数だけを使って表や箇条書きに含まれる表現間の意味的類似性を求められるため，HRAM より高速に単語意味クラスを獲得できる．

一方，表現間の共起関係を利用し，新聞記事などのコーパスから単語意味クラスを自動獲得する研究も多くある[1),3)~8),10)]．ここでは代表的なものについて触れる．

Church ら[1] は単語の出現頻度を基に 2 単語間の相互情報量を求めることで，$\langle doctors, nurses \rangle$ のような意味的に類似した単語の組を獲得している．しかし，相互情報量だけでは単語どうしが「関係している」ということしか分からないため，必ずしも意味的に類似している単語の組が獲得できるとは限らない．Church らは，$\langle doctor, bills \rangle$ のような，関係はしているが意味的に類似していない単語の組も獲得されたと報告している．本手法でも，相互情報量を，表や箇条書きか

ら抽出された表現どうしの意味的類似度の計算に用いている．そのため，$\langle doctor, bills \rangle$ のような関係はあるが意味的に類似していない表現どうしを「意味的に類似している」と判断してしまう恐れがある．しかしながら，表や箇条書きは意味的に類似した表現の分類に利用されやすいため，$\langle doctor, bills \rangle$ のような「関係はあるが意味的に類似していない」表現どうしが同一の表・箇条書きに含まれることは稀であると考えられる．本手法では，この**「表や箇条書きの要素である」**という制約を用いることで，*doctor* と *bills* を同時に含むような単語意味クラスが獲得されるのを最小限に抑えることを狙う．

Terra ら[10] は，相互情報量，χ^2 分布，対数尤度比，平均相互情報量などの統計量を，単語間の意味的類似度を測る際の尺度としてどの程度妥当か，という観点から比較している．Terra らのいう「意味的類似度」は 2 単語間のものであり，本手法で求めようとする，表現の集合の「単語意味クラスらしさ」とは異なる．Terra らが比較した各統計量は，Church らと同様に，表現どうしが「関係している」ということしか分からないため，「関係はあるが意味的に類似していない」表現どうしに対しても高いスコアを与えてしまう可能性がある．そのため，Terra らが比較した統計量を使って本手法で求めようする「単語意味クラスらしさ」を求める際は，「表現どうしが同一の表もしくは箇条書きの要素である」のような，統計的尺度とはまったく異なる別の手がかりが必要になる．

Turney[5] は，検索エンジンを使って 2 単語間の意味的類似度を計算する手法を提案している．具体的には，検索エンジンが提供する OR や NOT，近接などの検索オプションを用いて求めたヒット件数を基に 2 単語間の相互情報量を計算し，その値を意味的類似度としている．Turney によれば，検索オプションを用いずに得たヒット件数より，検索オプションを用いて求めたヒット件数から相互情報量を計算した方が，高い精度で意味的類似度を求められると報告されており，この結果は興味深い．Turney の手法では，2 単語間の意味的類似度を求めることを目的としており，表現の集合の単語意味クラスらしさを求めることを目的としている本手法とは，求めようとしている類似度の対象が異なる．また「ヒット件数を用いて相互情報量を計算する」という観点から両手法を比べると，Turney の手法では，ヒット件数を得る際に近接などの検索オプションが必要となるのに対し，本手法では検索オプションを必要としないため，本手法の方がより汎用的であると考えられる．

Riloff ら[6] や Roark ら[7] は，与えられた複数の表現（たとえば，*car* と *plane*）と意味的に類似している表現（*tank* や *helicopter* など）を，同格表現や並列名詞句を基に新聞記事などのコーパスから獲得する手法を提案している．本手法は同格表現などの代わりに，HTML 文書中の表・箇条書き，検索エンジンより得られるヒット件数を手がかりとして用いており，新聞記事や HTML 文書を構文解析しなくても，表現どうしの「共起の強さ」をとらえることができる．そのため，大量の HTML 文書から高速に単語意味クラスを獲得することが期待できる．

そのほかでは，単語の係り受け関係を用いて単語意味クラスを獲得する研究がある[3),4),8)]．Lin[3] や Pantel ら[4] の手法では，係り受け関係から単語間の意味的類似度を求め，互いに類似した単語どうしをまとめることで単語意味クラスを獲得している．また Rooth ら[8] は，係り受け関係と EM 法を用い，事前に設けた単語意味クラスへの単語の所属確率を推定し，単語意味クラスの獲得を行っている．本手法は，大量の構文解析済みコーパスが不要であるという点でどちらの手法とも異なる．

3. 提 案 手 法

3.1 概　　要

本手法では，以下に示す 2 つの仮説を用いる．仮説 1 は Shinzato らが提案したものである[9]．

仮説 1：HTML 文書中の表や箇条書きなどの構造には意味的に類似している表現が含まれやすい．

仮説 2：意味的に類似した表現どうしは文書中で共起しやすい．

この 2 つの仮説に従い，以下の 2 つのステップを経ることで単語意味クラスの獲得を行う．

ステップ 1：HTML 文書中で表や箇条書きなどの構造を使って分類されている自然言語表現の集合（以下，**関連表現集合**）の抽出

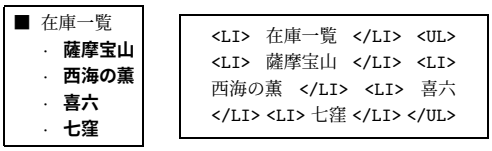
ステップ 2：Support Vector Machine（SVM）による関連表現集合の意味的類似性の判定

以下ではステップ 1，2 を順に説明する．

3.2 関連表現集合の抽出（ステップ 1）

ここでは，HTML 文書中の表や箇条書きに含まれる表現の集合を抽出する手法について簡単に述べる．この処理は HRAM のステップ 1 に相当している．より詳細な説明は該当論文[9] を参照されたい．

ステップ 1 では，仮説 1 に従い HTML 文書中に現れる各表現の持つ**パス**に注目することで，**関連表現集合**を抽出する．より具体的には，**同じパスを持つ表現**



(A) ブラウザ表示 (B) ソースコード

図 1 HTML 文書の例
Fig. 1 An example of HTML documents.

どうしをまとめ，それらを関連表現集合とする．ここでパスとは，HTML 文書中の表現を囲んでいるタグを，そのネストの順序に従ってリスト形式で表したものである．たとえば，**図 1** (A) に示した HTML 文書中の各表現は，同図 (B) のようにタグ付けされているため，それぞれ以下のようなパスを持っていると考えられる．

{(LI)，在庫一覧 }，{(UL，LI)，薩摩宝山 }，
{(UL，LI)，西海の薫 }，{(UL，LI)，喜六 }，
{(UL，LI)，七窪 }

そのため，この例の場合，ステップ 1 を適用することで，{ 薩摩宝山，西海の薫，喜六，七窪 } が関連表現集合として抽出される．関連表現集合は，特定の HTML タグに注目して抽出されるわけではないため，箇条書き以外の構造（たとえば，表やリストボックスなど）からも抽出可能である☆．

3.3 SVM による関連表現集合の意味的類似性の判定（ステップ 2）

ステップ 2 では，ステップ 1 で抽出した関連表現集合のうち，要素どうしが意味的に類似しているものを単語意味クラスとして獲得する．そのため，仮説 2 に基づき関連表現集合に含まれる表現どうしの共起の強さを求め，求まった共起の強さを手がかりにその集合が単語意味クラスと見なせるかどうか判定する．より具体的には，既存の検索エンジンより得られるヒット件数と，そのヒット件数より計算される相互情報量を共起の強さを表す指標として用いる．そして，それらの値を素性として SVM に与え，ステップ 1 で抽出された関連表現集合を SVM の決定関数の値に従って降順にソートし，その上位を単語意味クラスとして獲得する．

共起頻度および相互情報量を求めるため，まず，n 個の表現からなる関連表現集合から，表現の組を n 個生成する．具体的には，関連表現集合中の各表現 e について，e とは異なる表現 e' を無作為に選び出し，表現の組 $\langle e, e' \rangle$ を生成する．関連表現集合として { 薩摩宝山，西海の薫，喜六，七窪 } を考えた場合，たとえば，以下のような表現の組の集合が生成される．

{⟨ 薩摩宝山, 喜六 ⟩，⟨ 西海の薫, 薩摩宝山 ⟩，
⟨ 喜六, 七窪 ⟩，⟨ 七窪, 西海の薫 ⟩}

ここで注意したいのは，関連表現集合に含まれる表現の全組合せ（つまり，$n(n-1)/2$ 通り）について組を生成しない点である．仮説 2 に従うならば，関連表現集合に含まれる表現の全組合せを考慮し，後述する共起頻度と相互情報量を求める方が妥当であると考えられる．しかし，本研究では高速に単語意味クラスを獲得することが目的の 1 つであるため，n 個の表現を含む関連表現集合からは n 個の表現組しか生成しない．これにより，全組合せを考慮した場合は，相互情報量を求めるために $n + n(n-1)/2$ 回必要であった検索エンジンへの問合せが，$2n$ 回だけで済むようになり，大幅にその数を減らすことができる．これはつまり，高速に関連表現集合の意味的類似性の判定を行えることを意味している．全組合せを考慮しないことで，最終的な単語意味クラス獲得精度の低下が懸念されるが，我々の実験ではそれほど低下しないという結果が得られた．これについては，4.7 節で述べる．

次に，生成された各組について共起頻度と相互情報量を求める．$docs(x)$ を既存の検索エンジンより得られる表現 x のヒット件数，$docs(y, z)$ を表現 y と z を AND 検索したときのヒット件数とする．このとき，表現 e と e' の相互情報量 $I(e, e')$ を以下の式で計算する．

$$I(e, e') = \log_2 \frac{docs(e, e')/N}{(docs(e)/N) \times (docs(e')/N)}$$

本研究では，goo☆☆を利用して $docs(x)$，$docs(y, z)$ を求めている．また，N は検索エンジンが検索対象としている文書数であり，本研究では goo が検索対象としている HTML 文書の総数である 4.2×10^9 としている☆☆☆．

ここで次の関連表現集合 A，B を，意味的に類似しているものと，していないものに分類することを考えたい．

A { 薩摩宝山，西海の薫，喜六，七窪 }
B { インターナショナル，ご注文方法，ギフト券，トップセラー }

☆ 実験では，表の要素であることを表す <TD> タグや箇条書きの要素を表す <LI> タグのほかにも，他のページへのリンクを表す <A> タグ，文字色を変える <FONT> タグなどで囲まれている（より正確には，これらのタグ名をパスの最後に持つ）表現の集合も関連表現集合として獲得している．

☆☆ http://www.goo.ne.jp/

☆☆☆ http://help.goo.ne.jp/door/（2006 年 5 月現在）

表 1 相互情報量の例（$N = 4.2 \times 10^9$）
Table 1 Examples of pairwise mutual informaiton.

関連表現集合	e_i	$docs(e_i)$	e_j	$docs(e_j)$	$docs(e_i, e_j)$	$I(e_i, e_j)$
A	薩摩宝山	2.42×10^3	喜六	3.06×10^3	35	14.276918
	西海の薫	1.06×10^3	薩摩宝山	2.42×10^3	76	16.925030
	喜六	3.06×10^3	七窪	2.75×10^3	26	13.663650
	七窪	2.75×10^3	西海の薫	1.06×10^3	43	15.918942
B	インターナショナル	1.04×10^6	ギフト券	1.12×10^6	2.39×10^5	9.751174
	ご注文方法	8.43×10^5	インターナショナル	1.04×10^6	9.03×10^4	8.756855
	ギフト券	1.12×10^6	ご注文方法	8.43×10^5	7.87×10^4	8.451578
	トップセラー	1.29×10^5	ギフト券	1.12×10^6	7.95×10^4	11.174331

関連表現集合 A の各要素は芋焼酎の銘柄であり，意味的に類似していると考えられる．しかしその一方で，集合 B に含まれる要素は意味的に類似していない．この 2 つの関連表現集合について求めた共起頻度 $docs(e_i, e_j)$ と相互情報量 $I(e_i, e_j)$ の各値を**表 1** に示す．表より，関連表現集合 B に比べ A の各組について計算された相互情報量の値は全体的に大きいことが分かる．相互情報量の定義に従えば，表現 e と e' の出現が独立であると仮定した場合に，両表現の単独の出現確率 $P(e)$ と $P(e')$ から計算される共起確率 $P(e) \cdot P(e')$ よりも，実際に観測された共起確率 $P(e, e')$ の方が大きい場合にその値は 0 より大きくなる．そして，表現 e と e' の相互情報量の値が大きいほど，それらが共起しやすいことを意味している．このことから，関連表現集合 A に含まれる表現の方が，集合 B の表現より互いに強く文書中で共起するということが分かる．このことと，仮説 2 を考え合わせると，関連表現集合 A の方が B よりも，より意味的に類似した表現から構成されているといえる．そのため，これをうまくとらえられれば，ステップ 1 で抽出した関連表現集合を意味的に類似した表現からなるものと，そうでないものに分類できると期待される．このような考察に基づき，本手法では，共起頻度および相互情報量を素性として SVM に与え，ステップ 1 で抽出された関連表現集合の分類を行う．こうすることで，関連表現集合の分類を行う際，両方の指標を同時に考慮できるようになり，より高い精度での関連表現集合の分類が期待できる．

SVM に与えた素性の一覧を**表 2** に示す．実際は，共起頻度や相互情報量に加え，関連表現集合の要素数，集合に含まれる表現単独のヒット件数 $docs(e)$ なども素性として用いている．この理由は，関連表現集合が単語意味クラスかどうかを判別する際，これらも重要な手がかりになると考えたためである．また，相互情報量，共起頻度，ヒット件数に関しては，それぞれ最大値，2 番目に大きい値，最小値，2 番目に小さい値の 4 種類のみを素性として用いている．これはステップ 1 で獲得される関連表現集合の要素数 n が $4 \leq n \leq 30$ と一定でないため，求まったすべての値を同時に素性として利用することが難しいためである（このため，つねに値が存在することが保証されている上述した 4 種類の値だけを素性としている）．

表 2 本研究で用いている素性
Table 2 Features used in our procedure.

素性番号	説明
1	P 中で最も大きい相互情報量 $I(e, e')$
2	P 中で 2 番目に大きい相互情報量 $I(e, e')$
3	P 中で最も小さい相互情報量 $I(e, e')$
4	P 中で 2 番目に小さい相互情報量 $I(e, e')$
5	相互情報量 $I(e, e')$ の総和
6	相互情報量 $I(e, e')$ の平均
7	共起頻度 $docs(e, e')$ が 0 になる組数
8	共起頻度 $docs(e, e')$ の総和
9	共起頻度 $docs(e, e')$ の平均
10	P 中で最も大きい共起頻度 $docs(e, e')$
11	P 中で 2 番目に大きい共起頻度 $docs(e, e')$
12	P 中で最も小さい共起頻度 $docs(e, e')$
13	P 中で 2 番目に小さい共起頻度 $docs(e, e')$
14	C の要素数
15	文書頻度 $docs(e)$ が 0 の要素数
16	文書頻度 $docs(e)$ の平均
17	文書頻度 $docs(e)$ の総和
18	C 中で最も大きい文書頻度 $docs(e)$
19	C 中で 2 番目に大きい文書頻度 $docs(e)$
20	C 中で最も小さい文書頻度 $docs(e)$
21	C 中で 2 番目に小さい文書頻度 $docs(e)$

C：関連表現集合，P：C から生成された表現の組の集合

本研究ではより精度良く単語意味クラスを獲得するための工夫として，SVM を使って分類する際に得られる決定関数の値を，関連表現集合の**単語意味クラスらしさ**として解釈し，決定関数の値に従って各関連表現集合を降順にソートする．これにより，上位に順位付けられた関連表現集合だけを獲得することで，高い精度で単語意味クラスを得ることが可能になる．後の実験により，決定関数の値が関連表現集合の単語意味クラスらしさと正の相関があることを示す．

4. 評価実験

評価実験として，提案手法と類似手法である HRAM の比較実験，獲得された単語意味クラスの粒度の調査，用いた各素性の効果の確認，相互情報量の平均値および総和をソートの基準とするモデルとの比較実験，すべての組について共起頻度と相互情報量を求めた場合のモデルとの比較実験を行った．以下，本研究で設けた単語意味クラスの評価基準について述べた後，実験設定および各実験の結果について報告する．

4.1 評価基準

単語意味クラスを明確に定義することは難しい問題であるが，本研究では仮に以下の基準を満たす関連表現集合を単語意味クラスとした．

評価基準 A 関連表現集合中の **7 割以上の要素**に共通する具体的な上位語を考えることができれば，その集合を単語意味クラスとする．ただし，考えられる上位語として「物」や「事」などの一般的過ぎる語は除く．

本稿では「物」や「事」のような一般的過ぎるために上位語として適切でないと考えられる表現を**自明な上位語**と呼ぶ．自明な上位語は，意味的に類似していない表現の集合に対しても，各要素に共通する上位語と見なすことができるため問題である．たとえば，関連表現集合として { 自動車，机，人間，アイデア } が獲得されたとしよう．当然のことながら，常識的な観点からは各表現間に意味的な類似性を見ることはできない．しかしながら，仮に自明な上位語も上位語と見なすことを評価基準で許したとすると，得られた集合の各要素に共通する上位語として自明な上位語である「物事」を考えることが可能である．そのため，{ 自動車，机，人間，アイデア } は単語意味クラスと見なされることになり，これは我々の直感と反する．これはつまり，各要素に共通する上位語としてどのような表現を持つのかまでを考慮する必要があることを示唆している．以上の理由により，本研究では自明な上位語を上位語と見なさないようにした．

本研究では，日本語語彙体系[12] の一般名詞意味属性体系に含まれる各意味属性が，意味的な上位下位関係によって階層的に整理されていることを利用し，自明な上位語の獲得を行った．一般名詞意味属性体系では，137,966 個の表現（異なり数は 92,156）が 2,710 個の意味属性に従って分類されている．まず，一般名詞意味属性体系中の各表現について，ルートから何段目に位置する意味属性に含まれるのか調べた．このとき，複数の意味属性に属する表現については，階層の最も深い意味属性を，その表現が属する意味属性とした．ついで，上位 5 段までに位置する意味属性に属している 245 個の表現を，自明な上位語の候補として収集した．日本語語彙体系では，階層をルートから下に向かってたどった際，ルートからの段数と意味属性の抽象度の関係は，枝分かれごとにまちまちであり，意味属性ごとに枝分かれの数も異なっているため，階層構造の上位に位置する意味属性に属している表現が，必ずしも一般的過ぎる表現であるとは限らない．そのため，収集した各表現を 1 つずつ人手（具体的には著者のうち 1 人）で一般的過ぎるかどうかチェックした．その結果「個体」や「事象」など 154 個の表現が自明な上位語として得られた．本研究では，一般名詞意味属性体系中の意味属性に含まれている全表現の中から，収集された自明な上位語を除いた 92,002 個を，評価の際に上位語と見なす表現として抜き出した．以下，この抜き出した 92,002 個の表現からなるリストを**上位語リスト**と呼ぶ．

獲得された単語意味クラスの評価は次の手順で行った．まず，被験者に単語意味クラスとして獲得された関連表現集合を提示した．そして，関連表現集合の各要素に共通する上位語として，できる限り具体的な表現を思い浮かべてもらい，その表現が上位語リストに含まれているかどうかを評価ツールを使ってチェックしてもらった．このとき，想定された表現が上位語リストに含まれていれば，提示した関連表現集合の上位語として想定した表現を評価ツールに知らせ，次の関連表現集合の評価に移ってもらった．一方で，被験者によって上位語として想定された表現が上位語リストに含まれていない場合は，改めて異なる表現を想定してもらい，新しく想定された表現が上位語リストに含まれているかどうかを再びチェックしてもらった．この操作を何回か繰り返してもらい，どうしても各要素に共通する上位語を，上位語リスト中から見つけ出すことができない場合に限り，「上位語として適切な表現がない」ということを評価ツールに知らせてもらい，次の関連表現集合の評価に移ってもらった．

一般名詞意味属性体系の上位 5 段目までに位置する意味属性に，すべての自明な上位語が含まれているとは限らないため，上位語リストには自明な上位語が含まれていると考えられる．そのため，このことが上述の手順に従った評価の際に問題になると思われるかもしれない．しかしながら，後述する実験では，被験者により付与された上位語が階層構造中のどの位置に現れるかを確認しており，この実験によれば，被験者により付与された上位語のほとんどは，階層構造の末端に

位置する意味属性に含まれている語であった．この結果から，評価に用いた上位語リストに自明な上位語が含まれていても，さほど問題にならないと考えられる．

上述した評価基準 A は，考えられる上位語が単語意味クラスに含まれるすべての表現の上位語になっていなくてもよいため，比較的緩い基準のように思われる．そこで，本研究ではより厳しい評価基準として次の評価基準 B を設けた．

評価基準 B 関連表現集合中の**すべての要素**に対して共通する上位語を考えることができれば，その集合を単語意味クラスとする．ただし，自明な上位語は上位語と見なさない．

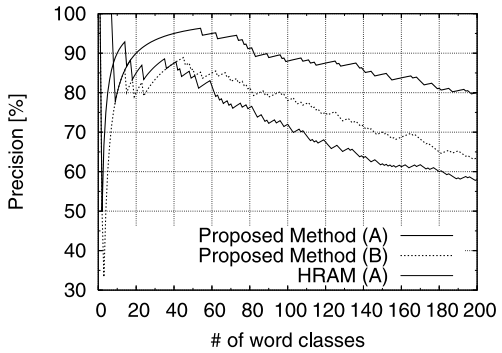
4.2 実験設定

実験にともない，1.0×10^6 件の HTML 文書（10.5 GB，タグ付き）を WWW より収集し，それらに対しステップ 1 を適用した．その結果，132,874 個の関連表現集合が得られた．そして，助詞を含む要素を持つ関連表現集合を削除した後，ランダムに 800 個（5,227 個の表現が含まれている）選び出し評価用データとした．助詞を含んでいる要素を持つ関連表現集合を削除したのは，意味的に一貫性を持った関連表現集合である可能性が低いと考えたためである．学習用データとしては，評価用の関連表現集合に含まれている表現を含まない関連表現集合を新しく 400 個（2,541 個の表現を含む）無作為に選び出した．そして，評価基準 A に従って単語意味クラスと見なせる/見なせないのラベルを，学習データとして選び出した関連表現集合に対して著者のうち 1 人が付与した．

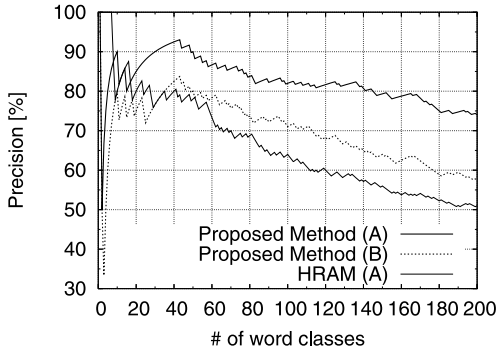
本研究では，*TinySVM* ☆ を用いて SVM の学習を行った．カーネル関数は，学習データを用いた予備実験で最も精度の高かった 2 次の *ANOVA* カーネルを用いている．この *ANOVA* カーネルを用いた場合の精度が最も高かった理由の理論的な分析は大変難しいが，他のカーネル関数，および他の次数の *ANOVA* カーネルに関しても予備実験を行った結果，収束しない，または高い性能が得られなかったため，本研究では 2 次の *ANOVA* カーネルを用いた．

4.3 提案手法と HRAM の比較実験

800 個の関連表現集合からなる評価用データに対して，提案手法および HRAM を適用し評価実験を行った．先述したように，提案手法により単語意味クラスと見なされた関連表現集合は，SVM の決定関数の値に従って降順にソートされている．その一方で HRAM の出力する上位語と関連表現集合の組も，その間の上位下位関係らしさを反映したスコアに従ってソートされている．そこで，両手法の出力する上位 200 個の関連表現集合を 4 人の被験者により評価した．もちろん，本来であれば，テストセットとして準備した 800 個すべてについて評価を行った方がよいと思われる．しかしながら，被験者への負担を考慮すると，800 個すべてについて評価してもらうことは難しく，この理由から本実験ではスコアの上位 200 個を評価の対象とした．評価結果を**図 2** に示す．図中 (A) は，提示した関連表現集合を 4 人の被験者のうち 3 人が単語意味クラスと判定した（具体的な上位語を付与した）場合の精度であり，(B) は 4 人全員が単語意味クラスと見なした場合の精度である．(A)，(B) ともに X 軸は提案手法，HRAM により単語意味クラスとして出力された関連表現集合の数を，Y 軸はそのときの精度（出力された関連表現集合のうち単語意味クラスと判断された集合の割合）を示している．図中の “`Proposed Method (A)`” と “`HRAM (A)`” は，提案手法および HRAM が出力した関連表現集合を評価基準 A に従って評価した場合の精度である．一方で “`Proposed Method (B)`” は，提案手法が出力した関連表現集合を評価基準 B に従って評価した場合の精度を示している．



(A) 被験者 4 人中 3 人が意味的に類似していると判断した場合



(B) 被験者 4 人全員が意味的に類似していると判断した場合

図 2 HRAM との比較実験
Fig. 2 Comparison with HRAM.

☆ `http://chasen.org/~taku/software/TinySVM/`

表 3 単語意味クラスとして獲得された関連表現集合と日本語語彙体系の比較
Table 3 Comparison between acquired semantic word classes and word classes included in the Nihongo-Goi-Taikei thesaurus.

要素の含まれ方	個数	関連表現集合の例
全要素が日本語語彙体系に含まれており，それらが同一の意味属性に含まれている	2	新宿（大字（その他），特別区，**駅名等**），渋谷（大字（その他），姓，特別区，**駅名等**），池袋（大字（その他），姓，**駅名等**），六本木（大字（その他），姓，**駅名等**），御茶ノ水（**駅名等**），駒込（大字（その他），姓，**駅名等**），下高井戸（大字（その他），**駅名等**）
全要素が日本語語彙体系に含まれているが，異なる意味属性に含まれている	4	雄勝町（町），稲川町（町，大字（町）），皆瀬村（村），東成瀬村（村）
一部の要素だけが日本語語彙体系に含まれている	19	力学（学問分野・学科），物理化学，数値解析，化学実験，無機化学（学問分野・学科），確率統計，化学応用，応用解析，物理学応用，微分積分応用
要素が 1 つも日本語語彙体系に含まれていない	90	古沢眼科，松山眼科，調布眼科医院，水野眼科クリニック，調布南口眼科医院

() 内はその単語を含む意味属性のラベル

図 2 のグラフより提案手法の方が HRAM に比べ高い精度で単語意味クラスを獲得できていることが分かる．図 2 (A) によれば，上位 80 個の関連表現集合（入力として与えた関連表現集合の 10%）を単語意味クラスとしたとき，提案手法の精度は，評価基準 A に従った場合で約 91%，基準 B では約 81%をそれぞれ示している．さらに，上位 200 個（入力の 25%）を単語意味クラスとした場合では，基準 A に従った場合で約 80%，基準 B で約 63%を示している．複数の被験者により行われた評価の一致度合いを示す *kappa* 統計量は，提案手法の出力した単語意味クラスの評価を行った場合が 0.69 であり，HRAM の場合は 0.78 であった．これらの値は論文 2) によれば *good* とされている値である．

ここまでで，提案手法により比較的高い精度で単語意味クラスを HTML 文書から獲得できることが分かった．次に検討すべきことは獲得された単語意味クラスの再現率であるが，まず 1 つの立場として WWW 上に膨大な量存在する HTML 文書に対して提案手法を適用することで，大量の単語意味クラスを獲得することが期待できるため，提案手法の再現率はさほど重要にならないと考えられる．また，もう 1 つの立場として，仮に再現率が重要であるとしても，現状では正解となる単語意味クラスが十分な量ないため，そもそも再現率を計算することが難しいという問題がある．ここでは 2 番目の問題をより詳細に検討するため，日本語語彙体系中の全意味属性を単語意味クラスのすべてとし，それらと提案手法により獲得された単語意味クラスとを比較した．具体的には，評価基準 B に従い，4 人の被験者全員が単語意味クラスと見なした 115 個の関連表現集合を対象に，それらが一般名詞，固有名詞の両意味属性体系に意味属性として含まれているかをみた．結果を**表 3** に示す．115 個の関連表現集合のうち，要素すべてが日本語語彙体系のいずれかの意味属性に含まれていたものは全部で 6 個あった．そのうち，2 個は要素がすべて同一の意味属性に含まれており，残りの 4 個は，要素が異なった意味属性に属していたものの，各要素が属するすべての意味属性の間には共通する親が存在するものであった．全要素が意味属性に含まれていた 6 個の関連表現集合以外の 109 個については，一部の要素だけが意味属性に含まれている集合が 19 個，すべての要素が意味属性に含まれていない集合が 90 個という結果であった．以上の結果は，日本語語彙体系にないものの被験者からして適切な単語意味クラスを提案手法では大量に獲得できていることを意味しており，なおかつ，既存のシソーラスなどではこの種の獲得手法の再現率を適切に評価することが難しいことを示している．

ついで，提案手法と HRAM が与えられた関連表現集合の意味的類似性を判定するために要する時間について考察する．提案手法は，表 2 に示した素性を用いて意味的類似性の判定を行っており，これらの素性を生成するためには，検索エンジンを利用して $docs(e)$ および $docs(e, e')$ を求めるだけでよい．そのため，検索エンジンへの問合せに要する時間を RT，関連表現集合の要素数を n とすると，$2nRT$ の時間で 1 関連表現集合の意味的類似性の判定を行うことができる．一方 HRAM では 1 要素ごとに，(1) 検索エンジンへの問合せ，(2) 検索結果上位 100 文書のダウンロード，HTML 文書中に含まれる文の (3) 形態素解析，(4) 係り受け解析の各処理を必要とする．そのため，1 文書あたり平均で m 文含まれているとすると，$(RT + 100DT + 100mMT + 100m\alpha PT) \times n$ だけの時間を，1 関連表現集合の意味的類似性の判定に要する．ここで，DT は 1HTML 文書をダウンロードするために要する時間，MT は 1 文を形態素解析する時間，PT は 1 文を係り受け解析する時間である．また，α は文中に関連表現集合中の要素が現れる確

表 4 獲得された単語意味クラスの例
Table 4 Examples of acquired semantic word classes.

順位	獲得された単語意味クラス	各被験者によって与えられた上位語			
		被験者 A	被験者 B	被験者 C	被験者 D
2	イラク問題関連，カネボウ関連，三菱自動車工業関連，りそな銀行関連，マイカルグループ関連，大木建設関連，森本組関連，マツヤデンキ関連，福助関連	N	N	N	関連
12	千葉聡子，中條浩介，趙成三，千草壽々子	医者	医者	氏名	人名
25	産業フェスティバル，正鬼様祭り，三川夏まつり，平等寺薬師堂大祭，観光キノコ園開園	N	催し	行事	行事
32	林寛子，高野ひろし，山口節生，あべ幸代，加藤盛雄，山根りゅうじ，あまたつ武夫，はやかわ忠孝，こみやま泰子，小川たくや，今澤まさかず，むらた文一	候補者	候補者	候補者	人名
45	オンワードスカイラークス，鹿島ディアーズ，オービックシーガルズ，富士通フロンティアーズ，日産スカイライナーズ，クラブハスキーズ，東京ガスクリエイターズ	チーム	アメフト	チーム	チーム
57	御枕屏風，和田嶺合戦図，八才竜女軸，泰嶺和尚書軸，西王母軸諏訪大社古図，慶応城下町図	絵画	絵画	絵画	絵画
60	光翼刃，雫刃，五光斬，桜華斬，盃割り	技	技	技	技
108	料理，ケーキ，サービス料，印刷物，挙式，飲物，引出物，装花，音響照明，席料，美容着付，介添料，控室料，新婦衣裳，新郎衣裳	N	N	N	N
114	樹脂，アルミニウム合金，ピアノ線，ウレタンゴム，ニトリルゴム，クロムモリブデン鋼	N	材料	材料	資材
122	旭丘小，第三福田小，第五福田小，連島西浦小	小学校	小学校	小学校	小学校
149	伊藤組，加藤建設，旭建設，佐野組，工藤組，齋藤組，千代田興業，藤和建設，中央土建，むつみ造園土木株式会社，英明工務店，山岡工業株式会社，加賀屋組，山二施設工業株式会社，秋田鋪道，三勇建設，株式会社長谷駒組，株式会社本郷建設工務所	土建屋	土建屋	土建屋	土木業
161	新人戦，インカレ，スチューデント，新人教育，関西選手権，インカレ予選，中部選手権，マリンカップ，インカレ団体戦，琵琶湖カップ，プレ新人戦，七夕カップ，あやめカップ，NT 選考	レース	大会	大会	大会
181	連結貸借対照表，連結損益計算書，海外事業，主要財務指標，単独貸借対照表，単独損益計算書，部門別売上動向，当期ハイライト，一株あたりデータ	N	N	N	N
187	一?三三，一八九?二〇六，二〇七?二二九，三五?一八八	N	N	号	N
190	山田正紀，三雲岳斗，森谷明子，典厩五郎	作家	著者	著者	小説家

N は適切な上位語が想定されなかったことを意味する．

率である．HRAM では関連表現集合に含まれる各要素の持つ係り受け関係だけが必要なため，それらを含む文のみを対象に係り受け解析を行っている．仮に，$m = 30$，$RT = 5\,\mathrm{sec}$，$DT = 1\,\mathrm{sec}$，$MT = 0.1\,\mathrm{msec}$，$PT = 1\,\mathrm{msec}$，$\alpha = 0.05$ と見積もると，HRAM は単語意味クラスかどうかの判別に $105.45n\,\mathrm{sec}$ 要することになる．それに対し，提案手法は $10n\,\mathrm{sec}$ で判別できるため，大雑把な見積りではあるが提案手法の方が高速に意味的類似性を判定できるといえる．

最後に，提案手法が単語意味クラスとして判断した関連表現集合の例を**表 4** に示す．単語意味クラスの評価は評価基準 B に従っている．表は左から，関連表現集合を SVM の決定関数が出力する値で降順にソートしたときの順位，獲得された単語意味クラス，4 人の被験者により付与された上位語の順に並んでいる．

4.4 獲得された単語意味クラスの「きめ細かさ」に関する調査

1 章で言及したように，本研究では「きめ細かい」単語意味クラスの獲得が目的の 1 つである．しかしながら，単語意味クラスの「きめ細かさ」を直接測ることは難しいため，ここでは，各単語意味クラスに対して被験者により付与された上位語の「具体さ」を利用して間接的に単語意味クラスの「きめ細かさ」を調査した．被験者にはあらかじめ可能な限り具体的な表現を上位語として単語意味クラスにつけてもらうよう指示しているため，「きめ細かい」単語意味クラスに対しては，詳細かつ具体的な表現が上位語としてつけられていることが期待できる．被験者により付与された上位語がどのくらい具体的かどうかは，日本語語彙体系の階層構造を利用して求めた．具体的には，上位語リストを獲得した一般名詞意味属性体系において，被験者により上位語として付与された表現が属する意味属性の下に，何段意味属性が登録されているか調べた．そして，下位に存在する意味属性の段数が少ない表現ほど具体的であると見なした．たとえば，**図 3** において，「伝導体」の段数を求めると，「伝導体」が属す意味属性「金属」の下には意味属性が 2 段存在する

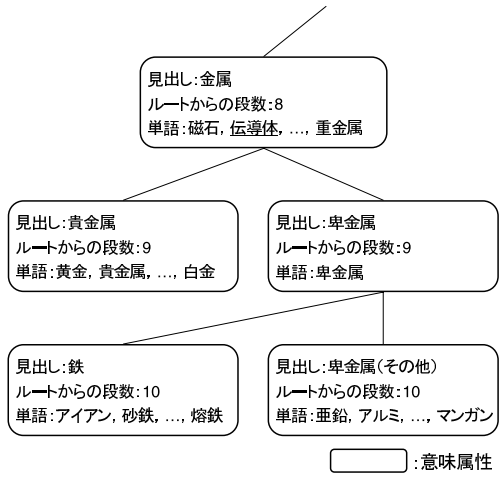


図 3 日本語語彙体系中の階層構造の例
Fig. 3 An example of the node hierarchy in the Nihongo-Goi-Taikei thesaurus.

表 5 付与された上位語の日本語語彙体系中での位置
Table 5 The positions of given hypernyms in the Nihongo-Goi-Taikei thesaurus.

	上位語を含んでいる意味属性の下に存在する階層数			
	0 段	1 段	2 段	3 段以上
被験者 A	95	4	9	7
被験者 B	79	21	12	3
被験者 C	91	9	9	6
被験者 D	99	8	6	2
平均	91（79.1%）	10.5（9.2%）	9（7.8%）	4.5（3.9%）

表 6 付与された上位語とその段数の例
Table 6 Examples of given hypernyms and their positions.

段数	付与された上位語の例
0 段目	選手，人名，サークル，休憩所，教職員，馬，学問，大学，ハーブ，牧師，駅，選挙区，ドライバー，チーム，講義，絵画，氏名，科目，班，研究者
1 段目	クラブ，ポジション，小学生，食材，先生，団体，通称，名，名前，料理，路線
2 段目	課，株式会社，官公庁，局，事業所，事務局，事務所，説，土建屋，部門
3 段目以上	企業，高山植物，植物，土地，組織

ため，“2” となる．日本語語彙体系のルートからの段数ではなく，下位に存在する意味属性の段数により表現の「具体さ」を求めた理由は，上位語リスト獲得の際にも述べたように，必ずしもルートからの段数がその意味属性の「具体さ」を表していないためである．4.3 節では，日本語語彙体系中の意味属性（単語意味クラス）と提案手法が単語意味クラスとして獲得した関連表現集合を比較して両者の重なり具合いについて調査したが，今回の実験では，日本語語彙体系と被験者より与えられた上位語を比較することで，提案手法が獲得する単語意味クラスの「きめ細かさ」について調査しているということに注意されたい．

4.3 節の実験で，提案手法が単語意味クラスとして出力した 200 個の関連表現集合のうち，評価基準 B に従い被験者 4 人全員が単語意味クラスと判断した 115 個について，付与された上位語の一般名詞意味属性体系中での位置を**表 5** に，被験者により与えられた上位語の例を**表 6** にそれぞれ示す．表 5 より被験者が付与した上位語の約 8 割が，意味属性体系の階層構造の末端に位置する（つまり段数 0）意味属性に含まれていることが分かる．一般名詞意味属性体系の末端に位置する意味属性には，「女優」や「馬」，「小学校」のような単語が全部で 87,984 個登録されており，提案手法が獲得した単語意味クラスの約 8 割は，それらを上位語に持っている．この結果から，提案手法により獲得された単語意味クラスの多くは，きめ細かい単語意味クラスであると考えられる．

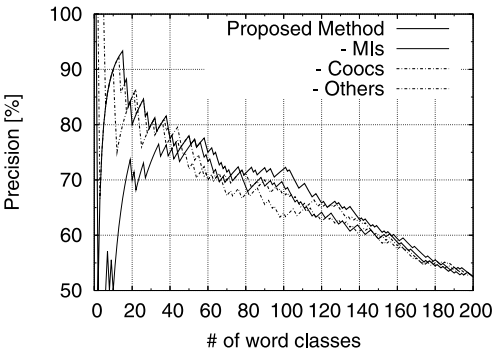


図 4 各素性集合の効果
Fig. 4 Contribution of each feature set.

上述の「きめ細かさ」に関する評価方法を，従来手法で獲得された単語意味クラスに対しても適用することで，提案手法と従来手法で獲得される単語意味クラスを「きめ細かさ」という観点から比較することができる．しかしながら，日本語を対象に自動構築された大規模な単語意味クラスは，少なくとも我々の知る限りでは存在しないため，両手法により獲得される単語意味クラスの「きめ細かさ」という点での評価は今後の課題としたい．

4.5 素性の効果

次に提案手法で用いている素性が，どの程度単語意味クラス獲得の精度向上に貢献しているのかを確認した．具体的には今回利用した素性（表 2）を以下の 3 グループに分け，各素性グループを用いなかったときに獲得精度がどの程度低下するかを見た．

MIs: $\{1, 2, 3, 4, 5, 6\}$,
Coocs: $\{7, 8, 9, 10, 11, 12, 13\}$,
Others: $\{14, 15, 16, 17, 18, 19, 20, 21\}$

上の各数字は表 2 にあげた素性番号と対応しており，*MIs* は相互情報量を用いた素性の集合，*Coocs* は共起頻度を用いた素性の集合，*Others* は相互情報量，共起頻度のどちらとも関係しない素性（たとえば，表現単体のヒット件数など）の集合をそれぞれ意味する．評価用データとしては，被験者への負担を軽減させるため，4.3 節の実験で入力として与えた 800 個の関連表現集合からランダムに 200 個選び出し，これらを用いた．これにより，入力として与えた関連表現集合全体にわたる提案手法の性能を大雑把にではあるが見ることができる．この実験では出力された関連表現集合を評価基準 A に従って評価し，4 人中 3 人の被験者によって上位語が想定された関連表現集合を単語意味クラスとした．

実験の結果を**図 4** に示す．図中の “Proposed

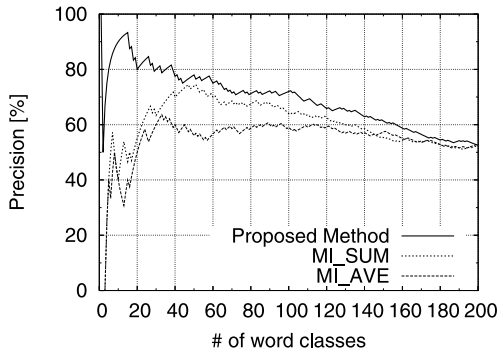


図 5 相互情報量の総和・平均値をソートの基準とした場合の性能
Fig. 5 Comparison with simpler methods.

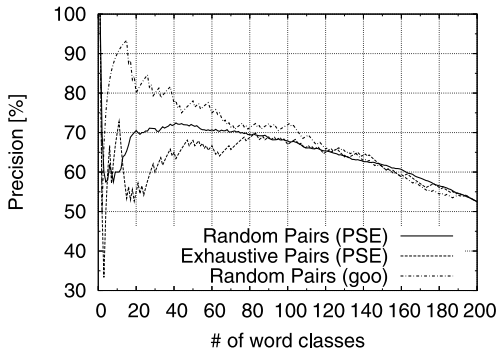


図 6 全組合せの共起頻度と相互情報量を求めた場合の性能
Fig. 6 Comparison with an exhaustive method.

Method" は提案手法の精度を，"-X" は提案手法から上にあげた素性集合 X を抜いた場合の精度を示している．提案手法の精度が最も高いことから，どの素性集合も精度の向上につながっていることが分かる．また，出力する関連表現集合の数を増やすと，獲得精度が徐々に低下していることから，SVM の出力する決定関数の値が関連表現集合の「単語意味クラスらしさ」をある程度とらえているということが分かる．

4.6 相互情報量の総和・平均値をソートの基準とするモデルとの比較

ついで，SVM を使って複数の素性を考慮することの有効性を検証した．具体的には，SVM を用いずに，相互情報量の総和および平均値で，各関連表現集合をソートするモデルとの比較実験を行った．実験に用いたデータは，4.5 節の実験と同じ関連表現集合である．つまり，最初にテストセットとした 800 個の関連表現集合から無作為に選び出した 200 個である．さらに評価基準も 4.5 節と同様に，獲得された関連表現集合を評価基準 A に従って評価し，4 人の被験者のうち 3 人が上位語を想定した集合を単語意味クラスとしている．

実験結果を**図 5** に示す．図中の 'MI_SUM' は相互情報量の総和で，'MI_AVE' は相互情報量の平均値で関連表現集合をソートした場合の精度をそれぞれ示している．図より単純に相互情報量の総和，または平均値に基づいてソートするよりも，提案手法の方が高い精度で単語意味クラスを獲得できていることが分かる．この結果から，SVM に複数の統計量を素性として与えることで，より適切なスコア付けが学習されたといえる．

4.7 無作為に n 組選択することの影響

提案手法では，高速に単語意味クラスを獲得するために，n 個の表現からなる関連表現集合に対して，3.3 節の手順で生成された n 組の表現組についてのみしか共起頻度および相互情報量を求めていない．そこで，すべての組（$n(n-1)/2$ 組）について共起頻度および相互情報量を求めた場合のモデルとの比較実験を行った．この実験では，検索エンジン goo を用いる代わりに，1.74×10^7 件の HTML 文書（191 GB，タグ付き）を検索対象とした全文検索エンジンを用いた．以下，この検索エンジンを *Private Search Engine*（PSE）と呼ぶ．goo を利用しなかった理由は，(1) 小規模な検索エンジンを用いた場合にどの程度の精度で単語意味クラスを獲得可能なのか確認したかった，(2) すべての組についてヒット件数を求めると，検索エンジンに対して多大な負荷をかけてしまうといった 2 つの理由による．検索エンジンを goo から PSE に変更したこと以外は，4.5 節および 4.6 節で行った実験と同じ設定で実験を行った．すなわち，最初にテストセットとして準備した 800 個の関連表現集合から無作為に選び出した 200 個を評価用データとし，評価基準 A に従って獲得された単語意味クラスの評価を行い，4 人の被験者のうち 3 人によって上位語が想定された関連表現集合を単語意味クラスとしている．

比較実験の結果を**図 6** に示す．図中の "Random Pairs (PSE)" は，PSE を用いた場合の提案手法の精度を，"Exhaustive Pairs (PSE)" は PSE を用いてすべての組合せについて共起頻度および相互情報量を求めた場合の精度をそれぞれ示している．また，"Random Pairs (goo)" は検索エンジンに goo を用いた場合の提案手法の精度である．つまり，この精度は図 4，図 5 における "Proposed Method" と同じである．なお，"Random Pairs (PSE)" については，偶然に高い精度が得られているということが考えられるため，10 回行った実験の平均を示している．

図より，関連表現集合に含まれる表現のすべての組合せについて共起頻度および相互情報量を計算しても，精度の向上がみられず，場所によっては精度が低下し

ていることが分かる．本実験では，提案手法が検索エンジン（PSE）に問合せを行った回数が 2,582 回であるのに対し，すべての組合せについて共起頻度を求めた場合は 5,714 回であった．このことから，提案手法は問合せ回数が半分以下なのにもかかわらず（つまり高速に動作する），すべての組合せを考慮した場合と同程度，場合によってはそれ以上の精度で単語意味クラスを獲得できていることが分かる．

また，図中の “Random Pairs (goo)” と “Random Pairs (PSE)” を比べることで，検索エンジンを goo から PSE に変えることにより精度が低下していることが分かる．上位 50 個（入力として与えた関連表現集合の 25%に相当する）の関連表現集合を出力した場合で，その精度の差は約 5%である．精度が低下した理由は，検索エンジン goo と PSE の検索対象としている文書数に大きな差があるためと考えられる．検索エンジン goo が対象としている文書数は 4.2×10^9 であるのに対し，我々が用意した PSE は 1.74×10^7 件の文書しか対象にしておらず，両者の差は 200 倍以上ある．そのため，PSE を用いた場合では goo を用いた場合よりも正しい表現間の共起関係を得ることができず，最終的な単語意味クラスの獲得精度が下がったのではないかと考えられる．しかしながら，200 倍という文書量の差を考慮すれば，5%程度の精度低下はさほど大きなものではないと思われる．

5. ま と め

本稿では，HTML 文書中に含まれる表・箇条書きなどの構造で分類されている自然言語表現の集合を，それらが意味的に類似しているかどうか判定することで，きめ細かい単語意味クラスを高速に獲得する手法を提案した．その特徴として，(1) 表・箇条書きに含まれる n 個の表現の意味的類似性を，構文解析などの重い処理を用いることなしに，$2n$ 回検索エンジンに問い合わせるだけで判定できる，(2) 既存の SVM 学習パッケージと商用検索エンジンを用いるだけで簡単に実装可能であるといったことがあげられる．

提案手法では，HTML 文書中から抽出された表・箇条書きなどの構造に含まれる自然言語表現の集合を，SVM の決定関数が出力する値に従ってソートし，その上位に順位付けされる集合を単語意味クラスとして獲得する．提案手法により獲得された単語意味クラスを 4 人の被験者により評価した結果，入力として与えた表現の集合の上位 10%を出力した場合，4 人中 3 人の被験者により，その 8 割が単語意味クラスと判定された．

今後の課題としては，獲得された単語意味クラス中に，意味的に類似していない表現が含まれることが実験により確認されたため，そのような表現を排除する手法の開発があげられる．このような手法が開発できれば，提案手法と組み合わせることで，より意味的に類似した単語意味クラスの獲得が期待できる．

謝辞 本研究を進めるにあたり，文部科学省科学研究費補助金（平成 15 年度若手研究（A）15680005，平成 15 年度萌芽研究 15650015）ならびに同省科学技術振興調整費（任期付若手研究員支援プログラム，新興分野人材養成プログラム）の支援を受けた．記して謝意を表する．

参 考 文 献

1) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Proc. 27th Annual Meeting of the Association for Computational Linguistics*, pp.76–83 (1989).
2) Landis, R. and Koch, G.: The measurement of observer agreement for categorical data, *Biometrics*, Vol.33, No.1, pp.159–174 (1977).
3) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp.768–774 (1998).
4) Pantel, P. and Lin, D.: Discovering Word Senses from Text, *Proc. ACM Conference on Knowledge Discovery and Data Mining* (*KDD-02*), pp.613–619 (2002).
5) Turney, P.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, *Proc. 12th European Conference on Machine Learning* (*ECML-2001*), pp.491–502 (2001).
6) Riloff, E. and Shepherd, J.: A Corpus-Based Approach for Building Semantic Lexicons, *Proc. 2nd Conference on Empirical Methods in Natural Language Processing* (*EMNLP-2*), pp.117–124 (1997).
7) Roark, B. and Charniak, E.: Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction, *Proc. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp.1110–1116 (1998).
8) Rooth, M., Riezler, S., Prescher, D., Carroll, G. and Beil, F.: Inducing a Semantically Annotated Lexicon via EM-Based Clustering, *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp.104–111 (1999).

9) Shinzato, K. and Torisawa, K.: Acquiring Hyponymy Relations from Web Documents, *Proc. Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, pp.73–80 (2004).
10) Terra, E. and Clarke, C.L.A.: Frequency estimates for statistical word similarity measures, *Proc. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp.165–172 (2003).
11) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer (1995).
12) 池原 悟，宮崎正弘，白井 諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林 良彦：日本語語彙体系 CD–ROM 版，岩波書店 (1999).
13) 河原大輔，黒橋禎夫：高性能計算環境を用いた Web からの大規模格フレーム構築，情報処理学会研究報告 2006-NL-171, pp.67–73 (2006).

(平成 18 年 7 月 10 日受付)
(平成 19 年 3 月 1 日採録)

新里 圭司

2002 年東京電機大学工学部情報通信工学科卒業．2004 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，2006 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．同年 10 月より京都大学大学院情報学研究科特任助教．自然言語処理の研究に従事．博士（情報科学）．

鳥澤健太郎（正会員）

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．自然言語処理，計算言語学の研究に従事．特に大規模テキストコーパスからの知識の自動獲得に興味を持つ．博士（理学）．