情報処理学会論文誌　データベース　Vol. 2　No. 2　1–9 (June 2009)

Webマイニングにおける語義曖昧性解消のための擬似負例を用いた能動学習

高山泰博[†1]　今村誠[†1]　鍛治伸裕[†2]
豊田正史[†2]　喜連川優[†2]

この論文では，多義語を含むWebの全文検索結果から，Webマイニングに用いる分析対象のデータ集合を高精度に抽出するための語義曖昧性解消（WSD）の研究について述べる．Webマイニングにおける WSDでは，学習に用いるラベル付き事例のうち，分析対象である正例は作成が容易であるが，分析対象以外の負例は予期せぬ意味が含まれるため作成が困難であるという課題があった．この論文では，負例の作成を容易にするために，正例のみで開始する能動学習の手法を提案する．提案手法の特長は，能動学習のプロセスにおいて，高い確信度で負例と予測されたラベルなし事例である“擬似負例”を負例として学習している点である．我々は，正例のみで負例のないWebデータに対する実験により，擬似負例を推定する提案手法が，人手により負例を作成した場合と比べて大差のないWSD精度を得ることを示す．

Active Learning with Pseudo Negative Examples for Word Sense Disambiguation in Web Mining

Yasuhiro Takayama,[†1] Makoto Imamura,[†1]
Nobuhiro Kaji,[†2] Masashi Toyoda[†2]
and Masaru Kitsuregawa[†2]

This paper studies a word sense disambiguation (WSD) to extract high accuracy dataset from a full text Web search result which contains polysemous words for web mining applications. The motivation of this study is the difficulty in the creation of the negative examples for learning, which include unexpected word senses especially in Web. In this paper, we present a method of active learning starting with only positive examples for WSD in order to facilitate the creation of the negative examples. The key feature is to learn with “pseudo negative” examples which have reliable confidence score as negative for the unlabeled examples during the active learning process. We show experimentally in the several Web data only with positive examples that our proposed method achieves close enough WSD accuracy to the method with the manually prepared negative examples.

1. はじめに

World Wide Webから有用な情報を発見することを目的としたWebマイニングでは，膨大なWebアーカイブからある検索要求に関連する文書を抽出する必要がある．たとえば，意見分析や評判分析[1)–3)]において信頼できる解析を行うためには，ある特定の製品，店舗，人物等に関する大量のテキストを抽出することが重要である．

Webアーカイブからテキストを検索する場合に，検索結果の中から分析対象として用いるテキストのみを取り出すことが必要になる．たとえば，雑貨店チェーンの“ロフト”の評判を分析する場合に，単純な全文検索の結果の中には，屋根裏部屋，ゴルフのクラブ面の角度，映画のタイトル，音楽のライブハウス等の，分析対象とは異なる意味の“ロフト”を含む無関係なテキストが含まれてしまう．このような意味の区別を扱うため，従来から自然言語処理において語義曖昧性解消（Word Sense Disambiguation，WSD）システムが用いられてきた．

WSDシステムの構築には，あらかじめ作成された大量の正例と負例のラベル付き訓練データを用いる．実際の応用では，検索要求に用いる検索語ごとに訓練データが準備されていないため，この訓練データをいかに作成するかが問題となる．

訓練データのうち，分析対象とするテキストの事例である正例の作成は容易である．たとえば，“ロフト AND 店”あるいは“ロフト AND 文房具”のように，上位語あるいは意味的に関連した語を用いて人手で検索要求を補強することによって，Webアーカイブから高適合率で（しかし，低再現率で）正例を検索できる．

一方，分析対象以外のテキストの事例である負例の作成は困難である．なぜなら，Webマイニングでは対象とする語の多くが固有名詞であり，人手で作成された辞書にはそれらの語の大部分の意味は載っていないからである．さらに，Webには不均一で多様な領域の文書が含まれるため，予期せぬ意味も見つかる．たとえば，この論文の共著者は誰も“ロフ

†1 三菱電機株式会社情報技術総合研究所
Information Technology R & D Center, Mitsubishi Electric Corporation
†2 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo

ト” にライブハウス，テレビの製品名に使われる “ベガ” にアトリエ，シャンプーの製品名に使われる “ツバキ” に競馬レースという意味があることを知らなかった．結果として，ラベルなし事例から負例を作成する際に，このような予期せぬ意味を見つけるのに時間を費やしてしまう．

この論文において，我々は，Web マイニングに適用する WSD システムを構築するために，正例とラベルなし事例だけから訓練データを作成する問題を扱う．ラベルなし事例から効率良く訓練データを作成する手法に能動学習がある．典型的な能動学習の手法では，正例と負例のラベル付き事例からラベルなし事例の確信度を計算し，最も確信度が低い事例に対して人の注釈者がラベルを割り当てる．そして，満足する分類結果が得られるまで，ラベルなし事例へのラベルの割当てを繰り返す．この論文の問題設定では，負例がない状態から学習を始めるため，正例と負例を必要とするこのような手法を用いることは困難である．我々は，この問題に取り組むために，正例とラベルなし事例で訓練した学習器による確信度により，ラベルなし事例の一部を負例であると見なした “擬似負例” を用いた能動学習の手法を提案する．

この論文の構成は下記のとおりである．2 章では，関連研究について述べる．3 章では，正例とラベルなし事例だけから開始する，WSD のための能動学習の手法を提案する．4 章では，提案手法を評価するための実験の設定について述べる．5 章では，実際に Web をクロールして収集した複数のデータに対する実験結果を示す．6 章では，この論文の結論を述べる．

2. 関 連 研 究

この論文では，最初の段階で負例がない場合に大量の負例の収集にコストがかかる Web マイニングのための WSD システムの構築における能動学習の問題を扱い，効率的な能動学習のために正例とラベルなし事例から推定した疑似負例を訓練データに用いる手法を提案する．この論文の関連研究には，WSD に関する研究と，訓練データ作成に関する研究がある．

曖昧な語の特定の文脈における意味を決定するための WSD のアプローチは，分類器の訓練の仕方に対応して，教師付き，教師なし，辞書ベースの 3 つの手法に分類される[4]．

教師付き WSD 手法は，ラベル付き訓練データ集合を基にして行われる．典型的な教師付き WSD 手法には，Bayes 分類器による Gale ら[5] の手法，情報理論に基づく Brown ら[6] による手法，および日本語単語の多義性解消に種々の教師付き学習手法を比較した村田ら[7] の研究がある．これらは，いずれもラベルなし事例を訓練データとして用いる手法を扱っていない．

教師なし WSD 手法には，Schütze[8] による Bayes 確率モデルに基づく EM アルゴリズムにより，与えられた個数のクラスタを構築する手法，Lin ら[9]，Pantel ら[10] のテキストからの概念発見の研究，新納ら[11] による EM アルゴリズムのループ回数の予測を用いた語義判別規則の学習の研究がある．これらの研究は，ラベル付き事例を訓練データに付け加えることにより精度を改善する機能はない．しかし，教師なし WSD は，ラベルなし事例だけで動作するので，ラベルなし事例を用いる我々の提案手法に対するベースラインアプローチと見なすことができる．

辞書ベースの WSD 手法は，白井ら[12]，Shirai ら[13] の研究のように辞書やシソーラス等の語彙的なリソースを用いる．Web マイニングでは対象とする語はほとんど固有名詞であり，それらの意味は人手で作成した辞書には記載されていないので，辞書ベースの WSD 手法はこの論文が対象としている，負例に予期せぬ意味がある場合には適合しない．

訓練データ作成に関する研究には，能動学習を用いるアプローチ，正例とラベルなし事例を用いた学習に関するアプローチがある．

訓練データ作成に能動学習を用いるアプローチには，Chan ら[14] による意味の事前確率の推定を含む領域適応化に関する研究と，Chen ら[15] や Zhu ら[16] による正例と負例がバランスしていない場合に能動学習を用いるテキスト分類の研究がある．これらは，いずれも最初の段階で負例がない場合を扱ってはいない．

正例とラベルなし事例を用いた学習に関するアプローチには，Liu ら[17]，Li ら[18]，Li ら[19]，Zhu[20] の研究がある．これらは，正例とラベルなし事例から負例を推定して訓練データを作成する手法を述べているが，能動学習を扱っていない．我々の知る限りでは，最初にまったく負例がない場合に正例とラベルなし事例から擬似負例を推定し，擬似負例を用いて分類器を訓練する能動学習に関する研究は行われていない．

3. 擬似負例を用いた能動学習

この章では我々の能動学習アルゴリズムを提案する．提案手法の大部分は，不確定性（uncertainty）サンプリングを用いた標準的な能動学習に従っている．すなわち，能動学習の各反復ステップにおいて，システムはラベルなし事例に対して確信度スコアを計算し，最も小さいスコアを持つ事例に対して人の注釈者が正あるいは負のラベルを割り当てる．提案アプローチの新規性は，負例を持たない状態から能動学習を始める点と，各反復において擬似負例を用いる点である．

3.1 分　類　器

この論文の実験では，教師付き曖昧性解消において典型的な naive Bayes 分類器を学習アルゴリズムとして用いる[4]．naive Bayes 分類器を用いた WSD の実行では，素性 f1, f2, . . ., fn を持つ事例 d に対して，次の式を最大化する意味 s を割り当てる．

$$\arg\max_{s} \; p(s|d) = p(s)\prod_{j=1}^{n} p(f_j|s) \tag{1}$$

意味 s は，Web マイニングの応用において対象とする意味であるときに正であり，そうでないときに負である．そこで，この論文では 2 値分類で語の曖昧性を解消する．ここで，正例は全文検索によって検索したものであり，負例は上述の擬似的なものである．なお，この論文では名詞のみを曖昧性解消の対象語として扱う．

この論文では，naive Bayes 分類器の訓練とテストにおいて，以下の素性を用いる．

(a) 文内の単語素性
対象語の周辺の $\pm n$ 文内に出現する内容語である．ここでは，$n = 1$，すなわち，曖昧性解消の対象語が出現する文を含んだ 3 文を用いる．この値は，実験により経験的に決定した．

(b) 文節内の先行単語素性
文節内において対象語の前に出現する内容語である．

(c) 文節内の後続単語素性
文節内において対象語の後に出現する内容語である．

(d) 係り文節素性
対象語が出現する文節に係る文節である．

(e) 受け文節素性
対象語が出現する文節を受ける文節である．

ここで，(a)，(b)，(c) の各単語素性は，その単語の基本形と品詞からなり，(d)，(e) の各文節素性は，文節内の単語と格マーカの対からなるものとする．naive Bayes 分類器の使用において，確信度スコア c(d, s) は，データ d の意味が s であると推定するものである．すなわち，データ d が素性 f1, f2, . . ., fn を持つとき，c(d, s) を以下の式によって計算する．

$$c(d, s) = \log p(s)\sum_{j=1}^{n} \log p(f_j|s) \tag{2}$$

```
01  # 定義
02    Γ (P, N):
03      正例 P，負例 N 上で訓練した WSD システム
04  # 入力
05    T ← ラベルなしデータを含む訓練データ集合
06  # 初期化
07    P ← T 上の全文検索による正の訓練データ集合
08    N ← φ (初期の負の訓練データ集合)
09    repeat
10        擬似負例 PN を生成する (図 2 参照)
11        Γ ← Γ (P, PN + N)：新しい WSD システム
12        # 能動学習による訓練データ集合の構築
13        c_min ← ∞
14        foreach d ∈ ( T - P - N ) do
15              WSD システム Γ により d を分類する
16              s(d) ← Γ による d に対する語の意味の予測
17              c(d, s(d)) ← d の確信度の予測
18              if c(d, s(d)) < c_min then
19                    c_min ← c(d), d_min ← d
20              end
21        end
22        人により d_min に正しい意味 s を与える
23        if s が正 then d_min を P に加える
24        else d_min を N に加える
25    until ラベル付き事例が所定の個数に到達
```

図 1　擬似負例を用いた能動学習
Fig. 1　Active learning with pseudo negative examples.

3.2 アルゴリズム

図 1 に提案アルゴリズムを示す．提案アルゴリズムの初期段階において，システムには，正例とラベルなし事例が与えられる．正例は，上位語あるいは意味的に関連した語を用いて

人手で検索要求を補強することにより全文検索で集める．

提案手法では，能動学習の各反復において，最初に擬似負例を生成する（図 1 行 10）．**図 2** に示すように，擬似負例の生成は，すべてのラベルなし事例を負例として naive Bayes 分類器を訓練し，その分類器によって各ラベルなし事例の語の意味を予測する．もし次の式 (3) により予測する確信度スコアが閾値 τ より大きければ，その事例は負である場合が多いため，その事例を擬似負例と見なすことにする（図 2 行 09〜11）．

$$c(d, psdNeg) = c(d, neg) - c(d, pos) \quad (3)$$

次に，与えられた正例と負例（擬似負例を含む）を用いて，再び各ラベルなし事例に対する確信度を計算する（図 1 行 12）．このとき，確信度の計算には naive Bayes 分類器に基づいた同じスコア関数を用いる．

能動学習による訓練データの構築において，Chan らの研究[14] と同様な不確定性サンプリング（図 1 行 13〜21）を用いる．このステップは，システム Γ によって一番確信度が低いと予測された最も不確か（uncertain）な事例 d_{min} を選択する．このとき，最も不確かな事例 d_{min} に対して，人が割り当てた正しい意味に従って，データ d を正のデータ集合 P あるいは負のデータ集合 N に付け加える（図 1 行 22〜24）．

Chen らの研究[15] と同様に，あらかじめ決めた個数の事例がラベル付けされるまで学習を繰り返す．

```
01        foreach d ∈ ( T - P - N ) do
02            WSD システム Γ (P, T-P) により d を分類する
03              c(d, pos) ← 式 (2) の定義により d が正
04                          と予測される確信度スコア
05              c(d, neg) ← 式 (2) の定義により d が負
06                          と予測される確信度スコア
07              c(d, psdNeg) = c(d, neg) - c(d, pos)
08              (d が擬似負例として予測される確信度)
09            PN ← d ∈ { ( T - P - N ) | s(d) = neg ∧
10                                  c(d, psdNeg) ≧ τ }
11                  (PN は擬似負例のデータ集合)
12        end
```

図 2　擬似負例の生成
Fig. 2　Generation of pseudo negative examples.

4. 実験の設定

この章では，提案手法を評価するための実験の設定について述べる．

4.1 実験データ

Web からクロールして収集した 4 種類の日本語の blog データを実験用のデータ集合として選択した．**表 1** に，実験に用いたデータの曖昧性を持つ語とそれぞれの曖昧な意味を示す．

表 2 に，曖昧な語，その語の意味の数，データ集合の数，データ集合における素性の数，正の意味を持つデータの割合を示す．

なお，データに対する正しいラベルの割当ては 1 人が行い，すべてのラベルの 48.5%については別の 1 人がチェックした．割り当てられたラベルの 2 人の間の一致度は 99.0%で

表 1　実験データ
Table 1　Experimental data.

語	正の意味	他の曖昧な意味
ベガ	製品名 （TV）	ラスベガス，サッカーチーム名， 愛称，星，馬，バカラのグラス， アトリエ，ワイン，ゲーム， 音楽関連会社名
ロフト	店舗名 （雑貨店チェーン）	屋根裏部屋，ゴルフクラブ面の 角度，音楽ライブハウス，映画
本田	人名 （サッカー選手）	人名（女優，音楽アーティスト， 他のサッカー選手）， ハードウェア店，自動車会社名
ツバキ	製品名 （シャンプー）	花の名前，着物，競馬レース， 花のツバキの成分，店の名前

表 2　評価のために選択したデータ
Table 2　Selected examples of data for evaluation.

語	意味の数	データ集合の数	素性の数	正の意味の割合
ベガ	11	5,372	164,617	31.1%
ロフト	5	1,582	38,491	39.4%
本田	25	2,100	65,687	21.2%
ツバキ	6	2,022	47,629	40.2%

あった．また，ラベルの割当てには，100 件あたり平均約 35 分の時間を要した．

4.2 実験条件

評価のために選択したデータの事例集合は，10%のテストデータ集合と 90%の訓練データ集合とにランダムに分割した．各々の曖昧な語に対する初期の正例は精度が 100%になるような厳密な検索要求で訓練データ集合から全文検索して構築した．初期状態では，訓練データ集合中の残りのデータはラベルなしである．**表 3** に，各全文検索要求の検索式，初期の正例の数，訓練データ集合における初期の正例の割合を示す．

実験は，負例なしの状態から開始する．能動学習プロセスの各反復において，事例データを 1 つずつ選択し，正例データ集合と負例データ集合に追加する．なお，評価のため，実験ではこのプロセスを初期の正例以外のすべての訓練データ集合中の事例が追加されるまで継続する．実験ごとに曖昧な語のデータ集合の数が異なるので，追加した事例の百分率によって，0 から 100%の 10%刻みで種々の図をプロットする．WSD の精度は，すべて負例の意味が必ずしも既知ではないというこの論文の WSD の設定に従って，各テストデータの推定結果が正か負であるかを計算することによって求める．

標準的な教師なし WSD である，EM アルゴリズムで学習した naive Bayes 分類器によるクラスタリングを学習アルゴリズムのベースラインとして提案手法と比較する．この実験では，クラスタリングの際のクラスタ数は，各データ集合に対して 2（正，負），および表 2 に示した意味の数の 2 種類を設定する．クラスタリングにはラベル付き事例は必要ないので，クラスタリングの対象データには訓練データとテストデータ両方のすべてのデータを使用する．

また，初期に負例がなくても提案手法による疑似負例を用いた適当な反復回数の能動学習により良い精度が得られることを示すために，人手で負例を用意した場合の実験を行う．具体的には，訓練データをサンプリングして人手で読み，表 3 の初期の正例と同数の負例を作成して初期の訓練データに用い，以降，追加する事例をすべて人手で読んだ訓練データを用いた naive Bayes 学習器による WSD の精度を求める．この精度を WSD の精度のほぼ上限と見なすことができる．

さらに，擬似負例の生成手法の良さを評価するために，正例以外をすべて負例にするという単純な方法で疑似負例を生成した場合，および学習プロセスの反復の際に乱数でランダムに事例を選択した場合をベースラインとして実験を行う．

表 3　評価のために選択したデータ
Table 3　Selected examples of data for evaluation.

語	初期の正例のための全文検索要求	正例データ集合の数（訓練データ集合中の割合）
ベガ	ベガ AND TV	316（6.5%）
ロフト	ロフト AND（雑貨 OR 文房具）	64（4.5%）
本田	本田 AND 圭祐	86（4.6%）
ツバキ	ツバキ AND 資生堂	380（20.9%）

5. 実験結果

この章では，実験の結果について議論する．

5.1 擬似負例の有効性

図 3 に，実験データ集合に対する各アプローチの WSD の精度の平均を示す．曲線 b-clustering，m-clustering は，それぞれクラスタ数として 2 および意味の数を指定した，クラスタリングを用いた教師なし WSD によるベースラインの学習アプローチの精度を表す．曲線 tentative は，正例以外をすべて負例にするという単純な方法で疑似負例を生成した場合のベースラインアプローチの精度を表す．曲線 random は，ランダムに事例を選択する場合のベースラインアプローチの精度を表す．曲線 uncertain は，図 1 で述べた不確定性サンプリングによる提案手法の精度を表す．また，曲線 human は，追加する事例をす

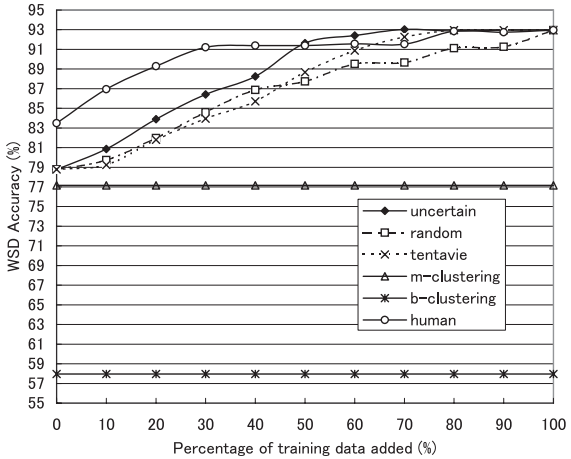


図 3　能動学習プロセスの平均
Fig. 3　Average active learning process.

べて人手で読んで負例を作成した場合の精度である．

提案アプローチの精度は，人手でラベル付けした事例を付け加える割合に従って徐々に増加する．たとえば，事例を 50%追加した時点における精度 uncertain はベースラインの学習アプローチ m-clustering よりも 14.6 ポイント高くなっている．これは，応用上の要求に合致して，ラベル付けがなされるほど精度が高くなることを示している．なお，クラスタリングの際のクラスタ数として 2 を設定した場合の精度の曲線 b-clustering は，クラスタ数として意味の数を設定した場合の曲線 m-clustering より 19.2 ポイントも低いため，以下の議論では，クラスタリングによるベースラインの学習アプローチについては曲線 m-clustering についてのみ言及する．

また，提案手法の曲線 uncertain は，すべてのベースラインよりも早く，精度の上限と見なした曲線 human に追いついている．これは，提案する擬似負例生成手法が他のベースラインよりも優れていることを示している．なお，曲線 uncertain は，曲線 human の初期の精度と同等な精度に追いつくのに約 20%の能動学習を要し，最終精度に近い 91%の精度に到達するのに約 50%（曲線 human では約 30%）の能動学習を要しているが，曲線 human では，初期の負例を作成するために，あらかじめ初期の正例以外の訓練データ中から平均 14.4%（ツバキの場合には，34.9%）のデータをサンプリングする必要がある．

図 3 の訓練データ追加のすべての割合で曲線 uncertain の精度は，曲線 random の精度よりも高くなっている．これは，提案する能動学習の不確定性サンプリングがランダムサンプリングよりも効率的であることを示している．同様に，図 3 の訓練データ追加のすべての割合で曲線 uncertain の精度は，曲線 tentative の精度よりも高くなっている．これは，提案手法による擬似負例の作成が，単純な方法による擬似負例の作成よりも効率的であることを示している．この結果は，能動学習プロセスにおいて，訓練データ集合にラベル付き事例として付け加えるべき事例を選択する際に「ラベルなし事例に対する信頼性の高い確信度スコアを得るために “擬似負例” の推定が有効である」という我々の主要な主張を立証している．また，応用上の観点から，この結果は，提案アプローチが「正しいラベルを割り当てるための作業コストを削減し，Web マイニングの現実の要求に対処するために，分析用の信頼性の高い基礎データを収集する時間を短くすることができる」ことを表している．

擬似負例だけを用いた提案手法のアプローチ，すなわち，訓練データに人手でラベル付けされた事例を何も付け加えない場合の精度は 78.8%である．一方，ベースライン学習アプローチ m-clustering の精度は，77.2%である．提案アプローチの初期の精度は，最初の正例を構築するための全文検索の検索要求に依存している．4 つのデータ集合に対する正例の検索要求が簡単に作りやすいように，多くの現実のアプリケーションにおいて，擬似負例だけを用いた提案アプローチの初期の精度は，ベースライン学習アプローチ m-clustering の精度とほとんど同じと見なすことができる．

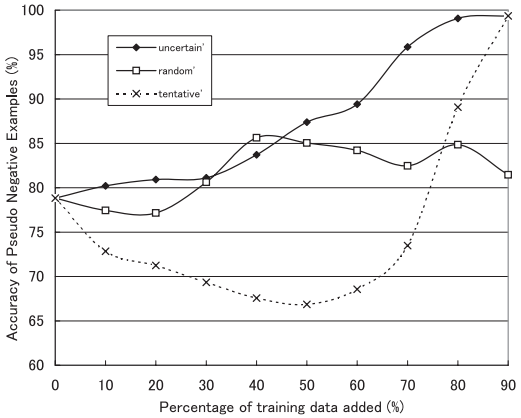


図 4　擬似負例の精度の平均
Fig. 4　Average accuracy of pseudo negative examples.

5.2　擬似負例の精度と WSD 精度との関係

擬似負例の精度の学習プロセスへの影響を調べるために，各反復における擬似負例の平均精度を**図 4** に示す．図 3 と同様に，反復数は追加した事例の百分率で表す．図 4 において，曲線 random′，uncertain′ および tentative′ は，それぞれ，ランダムに事例を選択した場合，不確定サンプリングで事例を選択した場合，正例以外をすべて負例にする単純な方法の場合における擬似負例の精度の平均を表す．図 4 において，提案手法による擬似負例の曲線 uncertain′ の精度が安定して上昇するのに対し，ランダムに事例を選択した場合の擬似負例の精度は不安定である．また，単純な方法による擬似負例の精度の曲線 tentative′ は，学習プロセスの後半に至るまでは他の 2 つの曲線 uncertain′，random′ と比べ低くなっている．

図 3 の学習プロセスの前半における WSD の精度は，曲線 uncertain，random，tentative の順であり，図 4 における擬似負例の精度の uncertain′，random′，tentative′ の順とほぼ対応する．これは，提案手法により作成した擬似負例の精度の良さが，WSD の精度に寄与していることを示している．なお，学習プロセスの途中から，必ずしも擬似負例の精度の順が WSD の精度の順と対応しないのは，訓練データの追加が進むにつれ人手判定により追

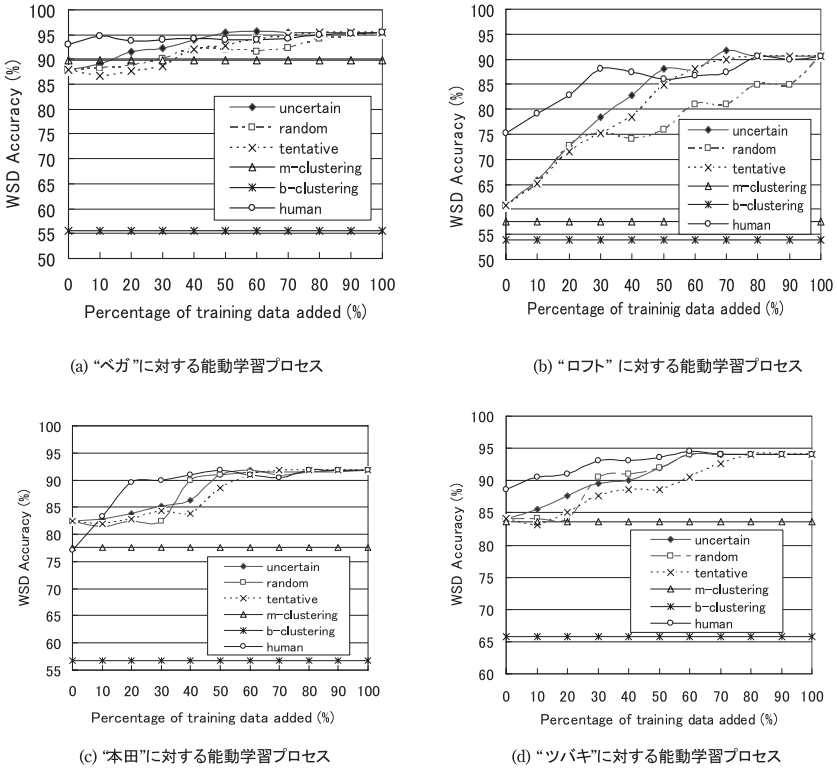


図 5　それぞれのデータ集合に対する能動学習プロセス
Fig. 5　Active learning process for each datasets.

加した負例の数が増えて，擬似負例の精度の影響が小さくなるためと考えられる．

5.3　データ集合への依存性

各データ集合に対する能動学習プロセスを**図 5** に示す．グラフ (a)，(b)，(c)，および (d) は，それぞれ表 2 の “ベガ”，“ロフト”，“本田”，および “ツバキ” のデータ集合に対する学習プロセスの精度を示している．図 5 の各グラフにおいて，曲線 b-clustering，m-clustering，tentative，random，uncertain および human が表すものはそれぞれ 5.1 節と同様である．

各実験データにおいて，曲線 uncertain の形は互いに類似しており，ラベル付き訓練データの追加に対し安定して精度が向上している．ラベル付き事例をあまり追加していない初期段階における曲線 uncertain の各精度は，それぞれ曲線 random および tentative の精度よりも高く，ラベル付き事例を追加するに従って，曲線 uncertain は他の 2 つの曲線より早い段階で曲線 human に追いついている．また，能動学習プロセスにおいて，提案手法の曲線 uncertain の精度は，ベースラインの曲線 m-clustering の精度よりもかなり高くなっている．この結果により，提案アプローチは多くの場合に適用可能なことが期待できる．さらに，人のラベル付けの作業コストを調整することによって，対象データの適合率と再現率の要求に対応できる．

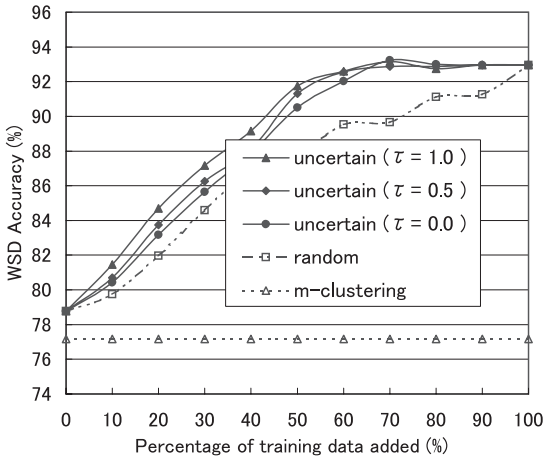


図 6　閾値 τ への依存性
Fig. 6　Dependency of threshold value τ.

グラフ (a) の “ベガ” とグラフ (b) の “ロフト” の学習プロセスで，曲線 uncertain と曲線 random の間にかなり違いがあるのに対して，グラフ (c) の “本田” に対する学習プロセスでは，曲線 uncertain と曲線 random の間に明確な違いがない．この理由は，“本田” のいくつかの負の意味の文脈が正の意味の文脈と類似しているからであると考えられる．具体的には，この実験において “本田” の正の意味はサッカー選手の名前であるが，負の意味にも何人かのサッカー選手の名前が含まれている．この振舞いの詳細な検討は将来の課題である．

5.4　閾値 τ への依存性

3.2 節で述べたように，我々は擬似負例を選択するための閾値として $\tau = 1.0$ を用いた．

この節では，τ の値を決定するための根拠となった実験について述べる．

図 6 は，擬似負例を選択するためのいくつかの閾値 τ による WSD の精度の平均を示したものである．図 6 において，曲線 m-clustering と random が表すものは 5.1 節と同じである．また，括弧付の曲線 uncertain は，括弧内の閾値 τ における提案アプローチの精度を表している．

図 6 では，$\tau = 1.0$ の場合が最も良い精度を示している．しかし，閾値 τ の間の精度の違いは提案アプローチとベースラインアプローチ m-clustering との違いに比べてとても小さい．これは，我々の提案アプローチが閾値 τ の選択には大きく依存していないことを示している．

6. おわりに

正例だけから開始する能動学習は，Web マイニングにおける WSD にとって重要である．この論文では，“擬似負例” を用いた能動学習の手法を提案した．提案手法の特長は，能動学習プロセスにおいて訓練データとして追加する事例を選択する際に，ラベルなし事例に対する信頼性の高い確信度を得るための擬似負例を推定することである．我々は，負例を持たない複数の日本語の blog のデータ集合に対して，標準的な教師なし曖昧性解消手法，ランダムに事例を選択した場合の能動学習による手法，正例以外のすべてを負例にするという単純な方法により擬似負例を作成した場合の能動学習による手法，および，追加する事例をすべて人手で読んできっちりとした負例を作成した場合と比較することによって，提案した WSD 手法の有効性を示した．

謝辞　本研究の一部は文部科学省リーディングプロジェクト e-Society 基盤ソフトウェアの総合開発「先進的なストレージ技術および Web 解析技術」による．実験データのクローリングを実施いただいた三菱電機（株）情報技術総合研究所の田村孝之氏，ならびに上記プロジェクトでの本研究にご協力いただいた皆様に，謹んで感謝の意を表する．

参 考 文 献

1) Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T.: Mining Product Reputations on the Web, *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.341–349 (2002).
2) Liu, B., Hu, M. and Cheng, J.: Opinion Observer: Analyzing and Comparing Opinions on the Web, *Proc. 14th International World Wide Web Conference on Data Mining*, pp.342–351 (2005).
3) Yi, J. and Niblack, W.: Sentiment Mining in WebFountain, *Proc. 21st International Conference on Data Engineering*, pp.1073–1083 (2005).
4) Manning, C. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, pp.235–239, MIT Press, Cambridge MA (1999).
5) Gale, W.A., Church, K.W. and Yarowsky, D.: A method for disambiguating word senses in a large corpus, *Computers and the Humanities*, Vol.26, No.5–6, pp.415–439 (1992).
6) Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and Mercer, R.L.: Wordsense disambiguation using statistical methods, *Proc. 29th Annual Meeting on Association for Computational Linguistics*, pp.264–270 (1991).
7) 村田真樹，内山将夫，内元清貴，馬　青，井佐原均：SENSEVAL2J 辞書タスクでの CRL の取り組み—日本語単語の多義性解消における種々の機械学習手法と素性の比較，自然言語処理，Vol.10, No.3, pp.115–133 (2003).
8) Schütze, H.: Automatic Word Sense Discrimination, *Computational Linguistics*, Vol.24, No.1, pp.97–124 (1998).
9) Lin, D. and Pantel, P.: Concept Discovery from Text, *Proc. 19th International Conference on Computational Linguistics*, pp.577–583 (2002).
10) Pantel, P. and Lin, D.: Discovering Word Senses from Text, *Proc. 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.613–619 (2002).
11) 新納浩幸，佐々木稔：EM アルゴリズムの最適ループ回数の予測を用いた語義判別規則の教師なし学習，情報処理学会論文誌，Vol.44, No.12, pp.3211–3220 (2003).
12) 白井清昭，八木恒和：辞書定義文を用いた低頻度語のための語義曖昧性解消モデルの学習，情報処理学会研究報告，自然言語処理研究会，2003-NL-20, pp.127–132 (2003).
13) Shirai, K. and Yagi, T.: Learning a Robust Word Sense Disambiguation Model using Hypernyms in Definition Sentences, *Proc. 20th International Conference on Computational Linguistics*, pp.917–923 (2004).
14) Chan, Y.S. and Ng, H.T.: Domain Adaptation with Active Learning for Word Sense Disambiguation, *Proc. 45th Annual Meeting of the Association of Computational Linguistics*, pp.49–56 (2007).
15) Chen, J., Schein, A., Ungar, L. and Palmer, M.: An Empirical Study of the Behavior of Active Learning for Word Sense Disambiguation, *Proc. main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp.120–127 (2006).
16) Zhu, J. and Hovy, E.: Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem, *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.783–790 (2007).
17) Liu, B., Dai, Y., Li, X., Lee, W.S. and Yu, P.S.: Building Text Classifiers Us-

ing Positive and Unlabeled Examples, *Proc. 3rd IEEE International Conference on Data Mining*, pp.179–187 (2003).
18) Li, X. and Liu, B.: Learning from Positive and Unlabeled Examples with Different Data Distributions, *Proc. 16th European Conference on Machine Learning*, pp.218–229 (2005).
19) Li, X., Liu, B. and Ng, S.: Learning to Identify Unexpected Instances in the Test Set, *Proc. 12th International Joint Conference on Artificial Intelligence*, pp.2802–2807 (2007).
20) Zhu, X.: A Semi-Supervised Learning Literature Survey, *Computer Sciences TR 1530*, University of Wisconsin (2007).

(平成 20 年 9 月 20 日受付)
(平成 21 年 2 月 17 日採録)

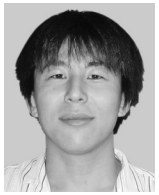
(担当編集委員　橋本 隆子)

高山　泰博（正会員）

1987 年 3 月九州大学大学院総合理工学研究科情報システム学専攻修士課程修了．同年 4 月三菱電機（株）入社．現在，同社情報技術総合研究所に勤務．自然言語処理，文書処理の研究開発に従事．1997〜1998 年スタンフォード大学言語・情報研究センター（CSLI）客員研究員．IEEE，言語処理学会，人工知能学会各会員．

今村　　誠（正会員）

1986 年 3 月京都大学大学院工学研究科数理工学専攻修了．同年 4 月三菱電機（株）入社．情報技術総合研究所勤務．博士（情報科学）．構造化文書処理，自然言語処理，CALS/EC システム等の研究開発に従事．情報処理学会 2006 年度論文賞，（社）日本電機工業会 2001 年度電機工業技術功績者発達賞等を受賞．電気学会等の会員．



鍛治　伸裕（正会員）

2005 年東京大学大学院情報理工学系研究科博士課程修了．情報理工学博士．現在，東京大学生産技術研究所特任助教．自然言語処理の研究に従事．

豊田　正史（正会員）

1994 年東京工業大学理学部情報科学科卒業．1999 年同大学大学院情報理工学研究科博士後期課程修了．博士（理学）．同年科学技術振興事業団計算科学技術研究員．2001 年東京大学生産技術研究所学術研究支援員．2004 年同大学特任助教授．2006 年より同大学准教授．ウェブマイニング，ユーザインタフェース，ビジュアルプログラミングに興味を持つ．ACM，IEEE CS，日本ソフトウェア科学会各会員．

喜連川　優（フェロー）

1978 年東京大学工学部電子工学科卒業．1983 年同大学大学院工学系研究科情報工学専攻博士課程修了．工学博士．同年同大学生産技術研究所講師．現在，同教授．2003 年より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，日本データベース学会理事，情報処理学会，電子情報通信学会各フェロー．ACM SIGMOD Japan Chapter Chair，電子情報通信学会データ工学研究専門委員会委員長歴任．VLDB Trustee（1997〜2002），IEEE ICDE，PAKDD，WAIM 等ステアリング委員．IEEE データ工学国際会議 Program Co-chair（1999），General Co-chair（2005）．本会副会長（2008〜2009）．