外れ値検出手法を利用した新語義の検出

新納　浩幸[†]・佐々木　稔[†]

本論文では対象単語の用例集合から，その単語の語義が新語義（辞書に未記載の語義）となっている用例を検出する手法を提案する．ここでのアプローチの基本は，新語義の用例が用例集合中の外れ値になると考え，データマイニング分野の外れ値検出の手法を利用することである．ただし外れ値検出のタスクは教師なしの枠組みになるが，新語義検出という本タスクの性質を考慮すると，一部のデータ（用例）にラベル（対象単語の語義）が付与されているという枠組みで考える方が適切である．そのため本論文では一部のデータにラベルがついているという教師付きの枠組みで外れ値検出を行う．具体的には 2 つの手法（教師付き LOF と生成モデル）を用い，それら出力の共通部分（積集合）を最終的な出力とする．この教師付き LOF と生成モデルの積集合を出力する手法を提案手法とする．実験では SemEval-2 日本語 WSD タスクのデータを用いて，提案手法の有効性を示した．また WSD のアプローチを単独で利用しただけでは，本タスクの解決が困難であることも示した．

キーワード：新語義，外れ値検出，LOF，生成モデル，One Class SVM，SemEval-2 日本語 WSD タスク

Detection of New Word Senses by the Outlier Detection Method

Hiroyuki Shinnou[†] and Minoru Sasaki[†]

In this paper, we propose a method to detect new word senses of a target word from sentences that contain it. To achieve this, we assume a new word sense sentence as an outlier of a data set constructed by sentences that contain the target word. Then using outlier detection methods in the data mining domain, we detect the new word senses. Generally, outlier detection methods are considered to be unsupervised. However, our method utilises data sets including some sentences with the labelled target word. Therefore, our outlier detection method is classified under the supervised framework. We propose an ensemble method of two methods to detect new word sense sentences: the supervised LOF (Local Outlier Factor) and the supervised generative model. The final output is the intersection of outputs of both methods. We demonstrate the effectiveness of our method using SemEval-2 Japanese WSD task data. Moreover we show that word sense disambiguation systems cannot solve our task by themselves.

Key Words: *new word sense, outlier detection, LOF, generative model, One Class SVM, SemEval-2 Japanese WSD task*

[†] 茨城大学工学部情報工学科, Department of Computer and Information Sciences, Ibaraki University

1 はじめに

本論文では対象単語の用例集合から，その単語の語義が新語義（辞書に未記載の語義）となっている用例を検出する手法を提案する．

新語義の検出は語義曖昧性解消の問題に対する訓練データを作成したり，辞書を構築する際に有用である．また新語義の検出は意味解析の精度を向上させる (Erk 2006)．また新語義の用例はしばしば書き誤りとなっているので，誤り検出としても利用できる．新語義検出は一般に Word Sense Disambiguation (WSD) の一種として行う方法，新語義の用例をクラスターとして集める Word Sense Induction (WSI) のアプローチで行う方法 (Denkowski 2009)，及び新語義の用例を用例集合中の外れ値とみなし，外れ値検出の手法を用いる方法 (Erk 2006) がある．ここでは外れ値検出の手法のアプローチを取る．ただしデータマイニングで用いられる外れ値検出の手法は教師なしであるが，本タスクの場合，少量の用例に語義のラベルが付いているという教師付きの枠組みで行う方が自然であり，ここでは教師付き外れ値検出の手法を提案する．

提案手法は 2 つの検出手法を組み合わせたものである．第 1 の手法は代表的な外れ値検出手法である Local Outlier Factor (LOF) (Breuning, Kriegel, Ng, and Sander 2000) を教師付きの枠組みに拡張したものである．第 2 の手法は，対象単語の用例（データ）の生成モデルを用いたものである．一般に外れ値検出はデータの生成モデルを構築することで解決できる．提案手法では第 1 の手法と第 2 の手法の出力の積集合を取ることで，最終の出力を行う．

提案手法の有効性を確認するために，SemEval-2 の 日本語 WSD タスク (Okumura, Shirai, Komiya, and Yokono 2010) のデータを利用した．従来の外れ値検出の手法と比較することで提案手法の有効性を示す．実験を通して，外れ値検出に教師データを利用する効果も確認する．また SVM による WSD の信頼度を利用した外れ値検出も行い，WSD システム単独では新語義の検出は困難であることも示す．

2 従来の新語義検出手法

2.1 WSD の信頼度の利用

WSD は語義を識別するタスクなので，WSD システムを利用すれば新語義を検出できると考えるのは自然である．WSD の対象単語 w の語義のクラスを C とする．関数 $f(x,c)$ はある WSD システムが出力する用例 x 中の w の語義が $c \in C$ となる信頼度とする．この WSD システムは $argmax_{c \in C} f(x,c)$ により語義を識別する．新語義の検出はある閾値 α を定め，

$$\forall c \in C \quad f(x,c) < \alpha \tag{1}$$

のときに x を新語義の用例と判定することで，新語義を検出できる．

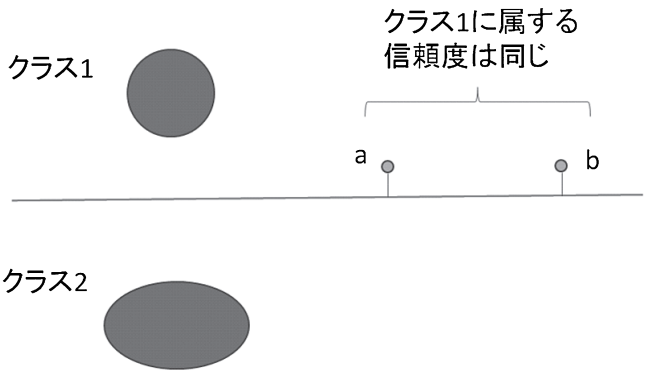

図 1 識別の信頼度と外れ値の度合い

ただし適切な α の値は単語毎に異なるはずであり，その設定は困難である．また WSD は識別のタスクであり，一般に WSD システムは SVM のような識別モデルをもとに構築される．そのためシステムは語義の識別精度が上がるように最適化されており，$f(x, c_i)$ の値は $f(x, c_j)$ との相対的なものである．つまり式 (1) により新語義が検出できる保証はない．例えば図 1 のような状況を考えてみる．図 1 のクラス 1 とクラス 2 を分離する直線が，分類器に対応する識別境界とする．データがクラスに属する信頼度は，一般に，識別境界までの距離で測るので，図 1 のデータ a とデータ b はクラス 1 と識別され，その信頼度は等しくなる．識別の場合はデータが識別境界のどちら側に属するかだけが重要なので，それで十分であるが，データ a とデータ b を比べると，明らかにデータ b の方がクラス 1 に属する信頼度が低い[1]．

2.2 WSI による検出

従来，新語義の検出は Word Sense Induction (WSI) というタスクの一部として行われてきた (Schütze 1998; Bordag 2006; 九岡，白井，中村 2008)．WSI は本質的には対象単語の用例を語義に基づいてクラスタリングするタスクである (Denkowski 2009)．用例集合中に新語義の用例があれば，それらも語義のクラスターとして出現するために新語義の検出として利用できる．

ただし陽に新語義を検出するには，得られたクラスターに語義のラベルを付与する必要がある (田中，中村，白井 2009)．Shirai は辞書に記述された語義の定義文を利用して，得られたクラスターに語義のラベルを付けることで新語義を検出しようとしている (Shirai and Nakamura 2010)．また Sugiyama は既存語義の用例を種用例として，用例集合を半教師なしクラスタリングによりクラスタリングした (Sugiyama and Okumura 2009)．種用例のないクラスターが新語

[1] 分類器が SVM の場合，データを高次元に写すので図 1 の例は特殊であるが，SVM でも同じ問題は内在する．また査読者から図 1 の例には問題があり削除するよう指示があったが，何が問題かが理解できなかったので，あえて本例は削除していない．

義のクラスターとなる．ただしどちらもクラスタリング自体の精度が悪く，新語義の検出までには至っていない．

本来，クラスターに語義のラベルを付けるためには，語義のラベル集合が必要である．語義のラベル集合を定めた場合に，WSI と WSD との違いはほとんどなくなる．WSD を行う前に教師なし学習であるクラスタリングを行うアプローチが，新語義の検出に有効かどうかは不明である．また用例を語義に基づいてクラスタリングする場合，クラスターの数の決め方が大きな問題になる (Agirre and Soroa 2007)．また新語義がクラスターを形成するという仮定は，多くの新語義に対して当てはまらない．クラスターを形成するくらいに，その語義の用例が存在するのであれば，その語義は新語義ではなく既に一般的な語義と考えられる．

2.3 外れ値検出による検出

新語義の用例を用例集合内の外れ値と見なし，外れ値検出の手法を利用して新語義を検出するアプローチがある．

Erk は外れ値検出手法の最近傍法を利用して新語義の検出を試みた (Erk 2006)．対象単語 w の語義が付与された用例集を D とし，用例 x の外れ値の度合い $out(x)$ を式 (2) で測り，この値が 1 以上の x を新語義の用例とした．ここで $d(x,y)$ は用例 x と用例 y 間の距離である．

$$out(x) = \frac{d(x,y)}{\min_{z \in D} d(y,z)} \qquad \text{where} \quad y = argmin_{y' \in D} d(x,y') \tag{2}$$

この式は D の中で x と最も距離が近いデータ y を選び，更にその y と最も距離が近い D 内のデータ z を選んで，$d(x,y)$ と $d(y,z)$ の比を取ったものである（図 2 参照）．

ただし最近傍法が妥当な精度を出すには，大量の訓練データを必要とするという問題がある．

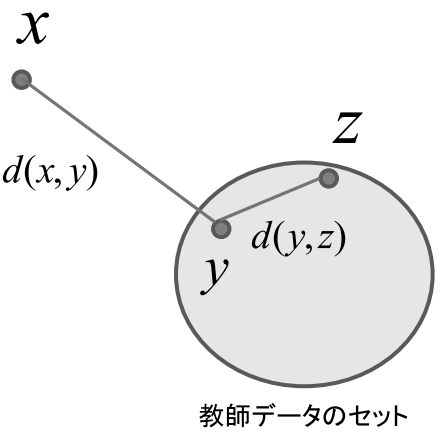

図 2 外れ値検出手法の最近傍法

3 外れ値検出手法

データマイニング分野の外れ値検出手法は非常に多岐にわたるが，その多くは変化点検出の手法に位置づけられる (山西 2009)．つまり時系列的にデータが生起するオンラインでのタスクに対する手法が中心である．新語義検出のようなバッチ的なタスクに対する手法としては，密度ベースの手法，One Class SVM，生成モデルによる手法が代表的な手法である．ここではこの 3 つの手法を本論文の提案手法との比較手法とする．

3.1 密度ベースの手法

外れ値検出は古典的にはマハラノビス距離を用いた距離ベースの手法が中心だが，それを改良したのが密度ベースの手法であり，密度ベースの代表的な手法が LOF である．LOF は，データの近傍の密度を利用することで，そのデータの外れ値の度合いを測り，その値によって外れ値を検出する．

LOF におけるデータ $x \in D$ における外れ値の度合いを $LOF(x)$ と表記する．ここで D はデータ全体の集合である．$LOF(x)$ を定義するために，いくつかの式を定義しておく．まず $kdist(x)$ は x に対する k 距離と呼ばれる値で，以下の条件を満たすデータ $o \in D$ との距離 $d(x, o)$ として定義される．

(1) 少なくとも k 個のデータ $o' \in D \setminus \{x\}$ に対して $d(x, o') \leq d(x, o)$ が成立する．

(2) 高々 $k-1$ 個のデータ $o' \in D \setminus \{x\}$ に対してのみ $d(x, o') < d(x, o)$ が成立する．

直感的には，上記のデータ o はデータ x からの k 番目に近いデータとなる．データ x から同じ距離を持つデータが複数存在する場合を考慮して，上記のようなテクニカルな定義になっている．

次に $kdist(x)$ を利用して，$N_k(x)$，$rd_k(x, y)$ 及び $lrd_k(x)$ を定義してゆく．

$$N_k(x) = \{y \in D \setminus \{x\} | d(x, y) \leq kdist(x)\}$$

これは x の k 近傍と呼ばれる集合であり，x との距離が $kdist(x)$ 以下になるようなデータの集合である．

$$rd_k(x, y) = \max\{d(x, y), kdist(y)\}$$

これは x から y への距離を表すが，x が y の k 近傍内に入る場合に，その距離を $kdist(y)$ で置き換えている．直感的にはデータ間の距離が近い場合に，k 距離で補正している．

$$lrd_k(x) = \frac{|N_k(x)|}{\sum_{y \in N_k(x)} rd_k(x, y)}.$$

これは x の k 近傍内のデータ y の $rd_k(x, y)$ の平均の逆数であり，これが x の密度を表して

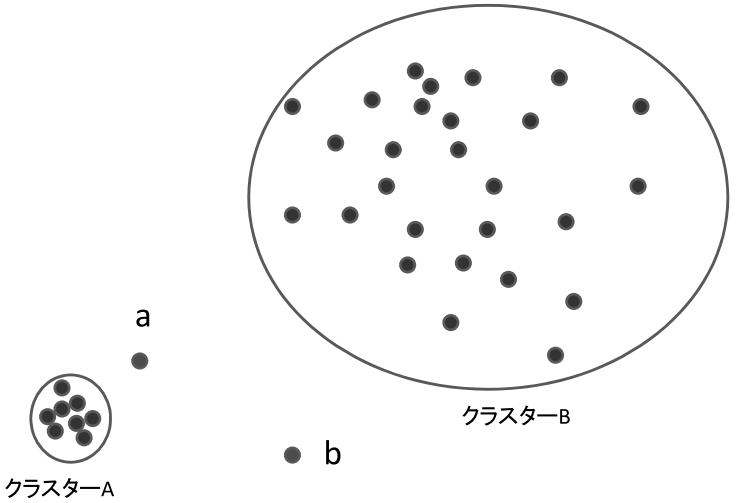

図 3 LOF による外れ値検出

いる．

これらの式を用いて，$LOF(x)$ は以下で定義される．

$$LOF(x) = \frac{1}{|N_k(x)|} \sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}$$

つまり x の k 近傍内のデータの密度と x の密度の比の平均を外れ値の度合いとしている．直感的には近くのデータの密度は高く，自身の密度が低い場合に外れ値の度合いが高くなる．また「近くのデータの密度は高く，自身の密度が低い」というのは，ある密度の高いクラスターがあり，そこから離れている独立のデータであるような場合である．

例えば図 3 では，データ a とデータ b が外れ値である．距離ベースの手法では，データ b は外れ値として検出できるが，データ a はクラスター A との距離が近いために検出できない．一方 LOF では，クラスター A の密度が高く，データ a の近辺にはデータがなく孤立しているので，外れ値として検出できる．

また LOF ではパラメータとして k が存在する．本論文では $k = 5$ としている．

3.2 One Class SVM

One Class SVM は ν−SVM (Schölkopf, Platt, Shawe-Taylor, Smola, and Williamson 2001) を利用した外れ値検出の手法である (赤穂 2008)．すべてのデータは +1 のクラスに属し，原点のみが −1 のクラスに属するとして，ν−SVM を使って 2 つのクラスを分離する超平面を求める．原点はすべての点に対して類似度が 0 となるために，外れ値とみなせる．また ν−SVM はソフ

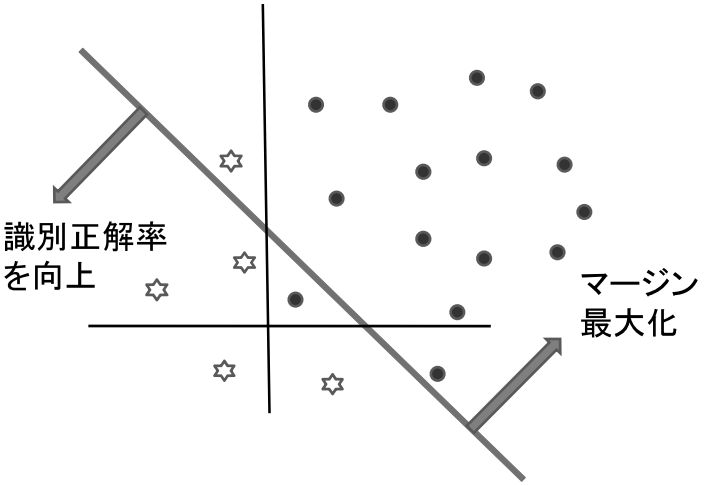

図 4 One Class SVM によるの外れ値検出

トマージンを利用するので，−1 のクラス側に属するデータを外れ値と判定する．

図 4 で概略を説明する．図の星形の点が外れ値である．原点は全ての点と内積が 0 となる，つまり類似度が 0 であるために外れ値と考える．図の星形の点（外れ値）も含め，原点以外のすべての点を正常値と考え，外れ値と正常値を分離する超平面を ν–SVM で求める．ν–SVM はソフト SVM であり，教師データのすべての点を正確に分離するわけではなく，少数の誤りを認める．図では原点付近に超平面（この場合，直線）を近づければ，識別の精度は向上するが，その場合，最大マージンが小さくなる．最大マージンを大きくしようとすると，識別の精度は下がる．このバランスをうまくとるような超平面を求めるのが ν–SVM である．最終的に原点側に属するデータが外れ値と判断される．

One Class SVM を利用する際には，用いるカーネル関数やどの程度のマージンの誤りを認めるかのパラメータの設定が結果に大きく影響する．本論文の実験では One Class SVM のプログラムとして LIBSVM[2] を用いた．カーネルは線形カーネルを利用し，マージンの誤りはパラメータ n に対応するが，$n = 0.02$ で固定した．

3.3 生成モデルによる手法

データ x の生成過程を確率モデル $P(x)$ でモデル化したものを生成モデルと呼ぶ．一般に潜在変数 z_i を導入し，ある確率モデル $P_i(x)$ の混合分布により $P(x)$ をモデル化する．

$$P(x) = \sum_i z_i P_i(x) \quad s.t. \quad 0 \le z_i \le 1, \ \sum_i z_i = 1$$

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

モデル化の後に，与えられたデータから EM 法などを利用して，z_i と $P_i(x)$ のパラメータを推定することで $P(x)$ を構成する．

データ x の外れ値の度合いとしては $-\log P(x)$ が用いられる．この値が大きいほど外れ値と見なせる．

4 提案手法

4.1 教師付き外れ値検出

一般に外れ値検出のタスクでは外れ値の定義が不可能である[3]．これは外れ値にラベルをつける意味がないことを示している．なぜなら仮にあるデータが外れ値であり，その外れ値にラベルをつけることができたとしても，他の外れ値がそのラベル付きの外れ値と類似している保証がないからである．また検出元となるデータ集合は，ほぼすべて正常値である．仮にデータにラベルをつけるとすれば，正常値のラベルだけになり，教師データに意味はない．これらのことから外れ値検出の手法は教師なしの枠組みにならざるをえない．

しかし新語義の用例を外れ値と見なした新語義検出のタスクの場合，一般の外れ値検出とは異なった 2 つの特徴がある．1 つは外れ値の定義が明確という点である．ここでの外れ値は新語義の用例であるが，新語義とは「辞書に記載されていない語義」というように明確に定義できる．もう 1 つは正常値のデータは語義のクラスターに分割されるという点である．しかもクラスターの数も明確である．一方，通常の外れ値検出では正常値の集合がクラスターに分割されるのか，されるとしてもいくつのクラスターに分割されるのかは不明である．

ここではこれらの特徴を利用して外れ値検出を行う．つまり，検出元となる対象単語の用例集の一部に，対象単語の語義のラベルを付与し，その設定のもとで外れ値検出を行う．

4.2 教師付き LOF

教師データを LOF で利用するには単純に教師データをテストデータに加えればよい．しかしその場合，教師データからも外れ値が検出される可能性がある．ここでは教師データを $k+1$ 倍してからテストデータに加えてデータセットを作り，そのデータセットに対して LOF を適用する．ただし k は LOF における $kdist$ で使われる k である．LOF の場合，教師データ x を $k+1$ 倍すると $kdist(x) = 0$ となり，教師データ x が外れ値として検出されることはなくなる．

教師データを $k+1$ 倍することで，テストデータに対して，外れ値検出の精度が高まるという保証はないが，いくつかの予備実験により経験的に精度が向上することは確認している．一般に教師データを増やせば検出の精度は高まる．また，教師データを増やせば既存の教師デー

[3] もし定義できるのであれば，その定義にあったデータを取り出せばよいだけなので，タスクとしての意味はなくなる．

タに対する密度が高まるはずなので，教師データを $k+1$ 倍することは精度を高める方向に作用する．また LOF は確率的な手法ではないので，明確には教師データの独立同一性分布を仮定していない．この点で同じデータを増やしても精度を落とす方向へ作用しないと考える．また注記として，教師なしの LOF も教師付き LOF も k の値が特に精度に影響を与えている．この点は考察の章で述べる．本論文では教師なしの LOF において $k=5$ としたが，教師付きの LOF でも $k=5$ とする．

4.3 教師データを利用した生成モデルの構築

対象単語 w の用例 x に対する生成モデル $P(x)$ を教師データを利用して構成する．w の語義を $z_i(i=1\sim K)$ としたとき，全確率の公式から以下が成立する．

$$P(x)=\sum_{i=1}^{K}P(z_i)P(x|z_i)$$

w の教師データが N 個あり，その中で語義 z_i のデータが n_i 個あれば，$\sum_{i=1}^{K}n_i=N$ であり，

$$P(z_i)=\frac{n_i}{N} \tag{3}$$

と推定できる．

問題は $P(x|z_i)$ の推定である．x は以下のような素性リストで表現されている．

$$x=\{f_1,f_2,\cdots,f_m\}$$

ここでは Naive Bayes で使われる素性間の独立性を仮定して，

$$P(x|z_i)\approx\prod_{j=1}^{m}P(f_j|z_i)$$

と近似する．教師データの中の語義が z_i となっているデータの中で f_j が出現した個数を $n(z_i,f_j)$ と書くことにする．このとき，

$$P(f_j|z_i)=\frac{n(z_i,f_j)}{n_i} \tag{4}$$

と推定できる．ただし式 (3) や式 (4) は頻度が 0 の部分があると不具合が生じる．そこで MAP 推定でスムージングを行い，以下の補正式を用いる (高村 2010).

$$P(z_i)=\frac{n_i+1}{N+K} \tag{5}$$

$$P(f_j|z_i)=\frac{n(z_i,f_j)+1}{n_i+2} \tag{6}$$

以上より $P(x)$ の値が求まる．外れ値の度合いは $-\log P(x)$ で測り，この値の大きなものを外

れ値の候補とする．

ここである閾値を定めて外れ値を検出することも考えられるが，単語毎に $-\log P(x)$ の値は大きく異なるために，固定した閾値を定めることはできない．そこでここでは単語毎に，検出対象のデータ（テストデータ）に対して $-\log P(x)$ を計算し，それらの値に対する平均 μ と分散 σ^2 を求める．$-\log P(x)$ の分布を正規分布と考え，以下の式の値（正規化した値）に対して閾値 θ を設けることにした．

$$\frac{-\log P(x) - \mu}{\sigma} \tag{7}$$

上記の正規化した値が θ 以上の x を外れ値とする．

ここでは予備実験を行い $\theta = 1.1$ とした．

4.4 教師付き LOF と生成モデルの積集合

本論文の提案手法は，前述した教師付き LOF による出力と，教師データを利用して構築した生成モデルの出力の共通部分（積集合）を出力するものである．

一般に外れ値検出のタスクは難しく，単一の手法ではなかなか高い検出能力が得られない．その 1 つの原因は誤検出が多いことである．提案手法の狙いは，異なったタイプの手法の出力の積集合を取ることで，誤検出を減らし，全体の検出能力を向上させることである．LOF と生成モデルは外れ値の捉え方が異なるために，出力の積集合を取る効果が期待できる．

5 実験

5.1 実験データ

SemEval-2 は語義曖昧性解消に関する評価型の国際会議であり，いくつかのタスクが設定されている．日本語 WSD はその中の 1 つである．通常の日本語の語義曖昧性解消のタスクであるが，最も特徴的な点は識別結果に新語義というカテゴリを含めている点である．つまりテストデータの中には設定された語義のどれでもないという答えがありえる．そのため，このタスクで用意された教師データとテストデータを用いることで，教師付きの枠組みでの新語義の検出手法の評価が可能である．

日本語 WSD の対象単語は 50 単語である．この中で「可能」「入る」は教師データ内に新語義の用例があるので，それらを外して残り 48 単語を実験対象とした．各単語を以下に示す．

名詞 21 単語
相手，意味，関係，技術，経済，現場，子供，時間，市場，社会，情報，手，電話，場合，
はじめ，場所，一，文化，ほか，前，もの

動詞 22 単語
会う，あげる，与える，生きる，入れる，教える，考える，勧める，する，出す，立つ，
出る，とる，乗る，始める，開く，見える，認める，見る，持つ，求める，やる

形容詞 5 単語
大きい，高い，強い，早い，良い

新語義は「意味」で 1 用例，「手」で 3 用例，「前」で 7 用例，「求める」で 1 用例，「あげる」で 2 用例，「はじめる」で 2 用例の計 16 用例存在する．これらが検出の正解となる．正解の用例を以下に示す．下線の単語が対象単語である．

1. …の開きが，ある意味で，科学技術と社会に…
2. …医業収益等は手入力…
3. …本部での集約も手入力，…
4. …経理コンピュータへの予算入力も手入力で…
5. …ランチ＝前十一時半～後 3 時．
6. …二十四日火，前十時～後 7 時…
7. …来年 3 月二十日木までの前十時～後十時，…
8. …ゆうゆうワイド（TBS ＝前8・三十）…
9. …三十日水までの前十一時半～後 2 時半，…
10. …，前十一時半～後 2 時半，…
11. …前十時半と後 6 時，本館 1 階正面口で…
12. …インフラ不安に要因を求め，その強化対策を…
13. …国を挙げて緑化を進めた．
14. …国をあげて緑化に取り組んだシンガポールは，…
15. 16. …「はじめる・はじまる」は「初」でなく，「始める・始まる」と書きます．

5.2 素性の設定

本手法を利用するためには，用例を素性ベクトルで表現しなくてはならない．そのために以下の 8 種類の素性を利用した．基本的 WSD で利用する素性である．なお対象単語の直前の単語を w_{-1}，直後の単語を w_1 としている．

e0 w の表記

e1 w の品詞

e2 w_{-1} の表記

e3 w_{-1} の品詞

e4 w_1 の表記

e5 w_1 の品詞

e6 w の前後 3 つまでの自立語の表記

e7 e6 の分類語彙表の番号の 4 桁と 5 桁

例えば以下は WSD の対象単語が 16 単語目の “経済” である文の形態素解析結果である[4].

```
<sentence>
<mor pos="名詞-固有名詞-組織名" rd="デンツー">電通</mor>
<mor pos="補助記号-読点" rd=", ">, </mor>
<mor pos="名詞-固有名詞-組織名" rd="ハクホー">博報</mor>
<mor pos="接尾辞-名詞的-一般" rd="ドー">堂</mor>
<mor pos="助詞-格助詞" rd="オ">を</mor>
<mor pos="名詞-普通名詞-副詞可能" rd="ハジメ">はじめ</mor>
<mor pos="名詞-普通名詞-一般" rd="ジョーイ">上位</mor>
<mor pos="名詞-数詞" rd="ゴ">五</mor>
<mor pos="接尾辞-名詞的-助数詞" rd="シャ">社</mor>
<mor pos="助詞-副助詞" rd="クライ">くらい</mor>
<mor pos="助動詞" rd="ナラ" bfm="ダ">なら</mor>
<mor pos="名詞-普通名詞-一般" rd="エイチピー">HP</mor>
<mor pos="助詞-格助詞" rd="オ">を</mor>
<mor pos="動詞-一般" rd="ツクル" bfm="ツクル" >作る</mor>
<mor pos="形状詞-一般" rd="ジンテキ">人的</mor>
<mor pos="名詞-普通名詞-一般" rd="ケーザイ" sense="X">経済</mor>
<mor pos="接尾辞-形状詞的" rd="テキ">的</mor>
<mor pos="名詞-普通名詞-一般" rd="ヨユー">余裕</mor>
<mor pos="助詞-係助詞" rd="モ">も</mor>
<mor pos="動詞-非自立可能" rd="アル" bfm="アル" >ある</mor>
<mor pos="助動詞" rd="デショー" bfm="デス">でしょう</mor>
<mor pos="助詞-接続助詞" rd="ガ">が</mor>
<mor pos="補助記号-読点" rd=", ">, </mor>
<mor pos="名詞-普通名詞-一般" rd="チューショー">中小</mor>
<mor pos="助詞-格助詞" rd="ノ">の</mor>
<mor pos="名詞-普通名詞-サ変可能" rd="ダイリ">代理</mor>
<mor pos="接尾辞-名詞的-一般" rd="テン">店</mor>
<mor pos="助詞-格助詞" rd="デ">で</mor>
<mor pos="助詞-係助詞" rd="ワ">は</mor>
```

[4] SemEval-2 の日本語 WSD タスクのデータはこの例のように，形態素解析済みのデータである．

```
<mor pos="連体詞" rd="ソンナ">そんな</mor>
<mor pos="名詞-普通名詞-一般" rd="ヨユー">余裕</mor>
<mor pos="助詞-係助詞" rd="ワ">は</mor>
<mor pos="動詞-非自立可能" rd="アリ" bfm="アル" >あり</mor>
<mor pos="助動詞" rd="マセ" bfm="マス">ませ</mor>
<mor pos="助動詞" rd="ン" bfm="ヌ">ん</mor>
<mor pos="補助記号-句点" rd="。">。</mor>
</sentence>
```

上記の用例から以下の素性リストが作成される．全体の素性リストが得られれば，全リストの各要素（素性）=（素性値）を各次元に設定することで，素性リストを素性ベクトル（実数値ベクトル）に変換できる．またここでは作成した素性ベクトルの大きさを 1 に正規化した．

e0=経済，e1=名詞-普通名詞-一般，e2=人的，e3=形状詞，
e4=的，e5=接尾辞，e6=人的，e6=作る，e6=HP，e6=余裕，
e6=ある，e6=中小，e7=2386，e7=1197，e7=11972

素性 e7 について注記しておく．上記例の場合，素性 e6 の値としては，「人的」「作る」「HP」「余裕」「ある」「中小」の 6 つ存在する．各々の分類語彙表の番号を調べると，以下のようになっている[5]．

作る ==> 2.386
余裕 ==> 1.1972
ある ==> 2.120 3.100

「人的」「HP」「中小」については分類語彙表に記載はない．「作る」の 2.386 から上位 4 桁を取り e7=2386 が作成される．また「余裕」の 1.1972 から上位 4 桁と 5 桁を取り e7=1197 と e7=11972 が作成される．最後に「ある」に関してだが，この単語からは素性 e7 は作成しない．本論文では全てひらがな文字からなる単語は多義語になっている場合が多い．そのため分類語彙表の番号を素性リストに含めてもノイズの方が多いと考え，そのような処理をしている．

5.3 実験結果

5.3.1 F 値による評価

まず F 値による評価実験の結果を表 1 に示す．LOF では LOF 値の大きなもの上位 3 つを取り出すことにする．3 つというのは，上位 1 つ，上位 2 つ，…，上位 5 つと各実験を行い，最も検出能力が高かった（F 値が高かった）ものである．OCS は One Class SVM の意味である．OCS ∩ LOF は LOF の出力と OCS の出力の積集合をとったものである (Shinnou and Sasaki 2010)．この 3 つが教師なしの外れ値検出に相当する．LOF-e は教師データを除いて LOF 値の

[5] 下位分類の番号は省略している．

表 1 実験結果（F 値）

手法	抽出数	正解数	F 値
LOF	144	2	0.0250
OCS	1228	3	0.0048
OCS ∩ LOF	53	0	0.0000
LOF-e	144	2	0.0250
OCS-e	618	3	0.0095
OCS ∩ LOF-e	50	0	0.0000
NN	377	6	0.0305
S-LOF	144	4	0.0500
G-model	307	5	0.0310
本手法	15	2	0.1290

高い上位 3 つをとったたもの，OCS-e は OCS の出力から教師データを除いたもの，OCS∩LOF-e は LOF-e と OCS-e の出力の積集合を取ったものである．また NN は (Erk 2006) で用いられた最近傍法であり式 (2) が 1.12 以上のものを取り出している．1.12 という閾値は出力結果から F 値が最も高くなるように設定した値である[6]．S-LOF は本論文で提案した教師付き LOF を指す．S-LOF では，LOF と同様，LOF 値の高い上位 3 つを取り出すことにする．また G-model は本論文で説明した生成モデルによるものである．この 6 つと S-LOF と G-model の出力の積集合を出力とする本手法の 7 つが教師付きの外れ値検出に相当する．

教師ラベルを全く使わない場合，教師データからも外れ値が検出されるので，F 値は低くなる．また単純に通常の検出を行った後に教師データを除く方法（表 1 の *-e の手法）よりも，積極的に教師データを利用した S-LOF の方が F 値が高い．

また S-LOF と G-model は検出の手法が異なるために，検出結果の重なりが少なく，結果的に両者の積集合を取る本手法の F 値が最も高かった．

5.3.2 平均適合率による評価

前節では手法の評価を F 値で行った．本節では全データに対して外れ値の度合いの順位を出力し，平均適合率を求めることで手法の評価を行う．

NN と G-model では出力の値（外れ値の度合い）をソートすることで，全データに対する外れ値の度合いの順位が得られる．LOF や S-LOF の場合は，単語毎に出力の値のスケールが異なるために，まず G-model で行ったような正規化を行い，単語毎の出力値のスケールを揃える．次に単語毎の出力値の上位 3 位までの出力値に 100 を加えた後に，全体をソートすることで，全データに対する外れ値の度合いの順位を得る．「上位 3 位までの出力値に 100 を加える」意

[6] 論文 (Erk 2006) で用いられている閾値は 1 である．

味は，単語毎の出力値の上位 3 位までを優先して出力することに対応する[7]．これは本来 LOF や S-LOF は単語毎に上位数件を外れ値として出力する手法であり，取り出さないデータの順位に意味があるかどうかは不明であるために導入した処理である．実際，この処理を行った方が，行わなかった場合よりも平均適合率は高かった．OCS の場合は，外れ値と判定したデータ群の重心を求め，その重心との距離によって，全データに対する外れ値の度合いの順位を得た．本手法 (G-model∩S-LOF) の場合，基本的に G-model の出力の値を外れ値の度合いとするが，本来の S-LOF における出力のデータ（単語毎の LOF 値の上位 3 件）に対しては，G-model での出力の値に 100 を加えた後に，全体をソートした．OCS∩LOF などの LOF 類と積集合を取る手法も本手法と同様に，LOF と組み合わせる方の手法のみで，まず外れ値の度合いを得て，次に本来の LOF における出力のデータ（単語毎の LOF 値の上位 3 件）に対して 100 を加えた後に全体をソートした．

実験の結果を表 2 に示す．表の 1 行目は手法名である．紙面の都合上 OCS∩LOF, OCS∩LOF-e, G-model は，それぞれ O∩L, O∩L-e, G-mdl と略記している．表の 1 列目は新語義の現れた個数，表内の数値はその個数の時点での適合率である．例えば，本手法の場合，1 番目に新語義の現れた順位は 7 であり，その時点での適合率は $1/7 = 0.14286$ であり，2 番目に新語義の現れた

表 2 実験結果（平均適合率）

	LOF	OCS	O∩L	LOF-e	OCS-e	O∩L-e	NN	S-LOF	G-mdl	本手法
1	.00926	.00465	.00709	.01587	.00952	.00714	.00847	.00990	.02174	.14286
2	.01449	.00873	.01399	.01887	.01739	.01389	.01124	.01869	.01515	.15385
3	.00820	.01181	.00845	.01415	.02362	.01230	.01657	.02752	.01648	.07143
4	.00541	.00156	.01084	.00978	.00318	.01575	.01270	.02817	.01476	.03960
5	.00466	.00194	.01272	.00855	.00396	.01887	.01362	.02242	.01818	.02632
6	.00417	.00228	.00225	.00776	.00465	.00441	.01630	.00402	.01217	.01899
7	.00249	.00264	.00261	.00505	.00538	.00513	.01703	.00407	.01365	.01737
8	.00279	.00282	.00292	.00563	.00576	.00575	.01168	.00439	.01504	.01320
9	.00314	.00308	.00327	.00633	.00629	.00641	.01230	.00407	.01355	.01440
10	.00347	.00301	.00340	.00700	.00611	.00673	.00984	.00452	.01167	.01558
11	.00381	.00306	.00364	.00769	.00618	.00721	.00980	.00497	.01175	.01160
12	.00413	.00326	.00353	.00833	.00658	.00696	.01023	.00542	.01233	.01175
13	.00426	.00317	.00354	.00855	.00634	.00699	.01072	.00586	.01151	.01231
14	.00372	.00310	.00373	.00743	.00619	.00737	.01101	.00630	.01043	.01160
15	.00373	.00321	.00361	.00744	.00631	.00711	.01037	.00643	.01010	.01068
16	.00343	.00340	.00352	.00673	.00667	.00699	.00918	.00679	.00964	.01041
AP	.00507	.00386	.00557	.00907	.00776	.00869	.01194	.01022	.01363	.03637

[7] 理論的には順位 1 位の出力値 + 1 で十分だが，本論文で扱う手法は全て順位 1 位の出力値が 100 よりもかなり小さいので，100 を加えるという単純な処理にしている．

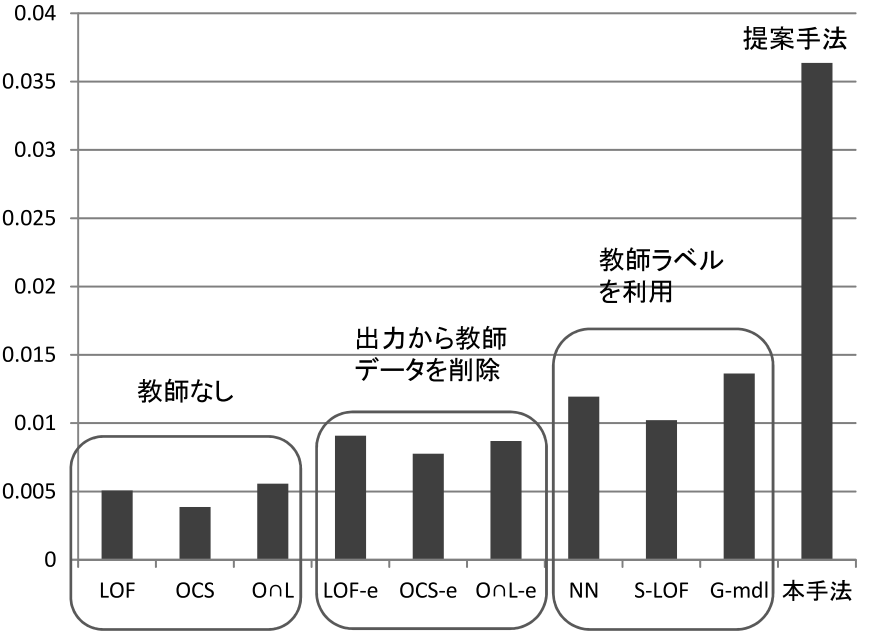

図 5 平均適合率

順位は 13 であり，その時点での適合率は $2/13 = 0.15385$ である．これを全ての新語義の個数 16 個まで調べ，各適合率の平均が平均適合率 (Average Presision = AP) である．そして各手法に対する平均適合率をグラフ化したものが図 5 である．

表 2 及び図 5 より，本手法が最も平均適合率が高いことが確認できる．また教師なしの手法にあたる LOF, OCS, OCS ∩ LOF の 3 つは 0.005 前後の値となり，教師ラベルを使わずに単純に出力結果から教師データを除く手法 LOF-e, OCS-e, OCS ∩ LOF-e の 3 つは 0.010 弱の値になり，教師付き外れ値検出手法にあたる NN, S-LOF, G-model の 3 つは 0.010 強の値になる．これらのことから教師データを外れ値検出に積極的に利用する効果も確認できる．

6 考察

6.1 教師データを $k+1$ 倍する効果

S-LOF は教師データを $k+1$ 倍した LOF であるが，この倍率を 1 から $k+1$ まで変化させた結果を表 3 に示す．なお倍率 1 倍は通常の LOF である．正解の検出数及び教師データからの検出（誤検出）は k 倍まではほぼ変化ないが，$k+1$ 倍することで急激に改善される．これにより教師データを $k+1$ 倍する効果が確認できる．

6.2 WSD による新語義検出

WSD の教師データが利用できるのであれば，WSD の分類器を学習し，その識別の信頼度を利用して新語義が検出できると考えるのは自然である．ただし単純にそのアプローチだけでは新語義の検出は困難である．

前述した素性を使い SVM を学習し，SemEval-2 日本語 WSD タスクのテストデータ 50 単語全てを対象に語義の曖昧性解消を行ったところ，平均正解率は 0.7664 であった．上記タスクの参加システム中最高の正解率は RALI-2 の 0.7636 であり (Okumura et al. 2010)，ここで学習できた SVM は十分能力が高いことがわかる．上記 SVM の学習には LIBSVM を用いたが，そこでは`-b`のオプションで識別の信頼度（その語義に属する確率値）を求めることができる．このオプションを用いて，閾値 θ 以下の信頼度のときに，その用例を新語義の用例とすることで新語義の検出を試みた．

閾値 θ の設定であるが，まず単純に 0.51 から 0.99 までの値を 0.01 刻みで設定し，その値を用いた場合の検出結果に対する F 値を求めた．そのグラフを図 6 に示す．$\theta = 0.73$ のときに検出数 388 正解数 4 となり F 値 が最大の 0.0198 を取る．また語義数が K の場合，SVM が出力する識別の信頼度は明らかに $1/K$ 以上の値になるので，語義数の影響を受けている可能性があ

表 3 S-LOF における教師データの倍率の変化

倍率	抽出数	抽出データ内の教師データ数	正解数
1	144	63	2
2	144	83	1
3	144	80	1
4	144	75	1
5 ($= k$)	144	82	1
6 ($= k + 1$)	144	0	4

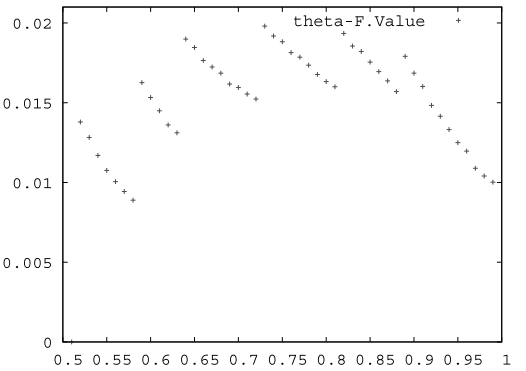

図 6 閾値と F 値

る．そこで閾値を $\theta = (1+\alpha)/K$ と設定し，α を 0.01 刻みで 0.99 まで試したときのグラフを図 7 に示す．$\alpha = 0.17$ のときに検出数 39 正解数 2 となり F 値 が最大の 0.0727 を取る．

F 値 0.0727 は表 1 で示された外れ値検出手法と比較すると，それほど悪いとも言えないが，WSD システム単独では新語義の検出が困難であることがわかる．

また平均適合率の評価も行っておく．システムが識別した語義の信頼度によって，全体のデータを（昇順に）ソートすることで，平均適合率を調べたところ 0.00638 となった．この値は表 2 に示した外れ値検出手法による平均適合率と比べると高い値とは言えない．平均適合率の観点からも，WSD システム単独では新語義の検出が困難であることがわかる．

上記では語義の識別の信頼度により新語義を検出するアプローチであったが，ここでは SVM を利用しているので，one-vs-rest 法を利用して，語義毎に SVM を学習し，すべての語義について否と判定されたものを新語義とするアプローチも考えられる．このアプローチによる評価も行っておく．

語義毎に SVM を学習する際にも LIBSVM の`-b`のオプションを用いる．語義毎の各 SVM が否と識別した信頼度を集め，その最小値 γ をそのデータの新語義の度合いとする．γ が閾値 θ よりも大きい場合に，新語義と判定する．$\theta = 0.5$ は語義毎の SVM 全てが否と判定したものを新語義と判定することを意味する．出力結果の分析から $\theta = 0.6996$ のときに検出数 33 正解数 1 となり F 値 が最大の 0.0408 を取る．また one-vs-rest 法を利用した場合の平均適合率も調べた．γ の値を新語義の度合いとし，全データに対して新語義の度合いの順位を出力することで平均適合率が求まる．結果，平均適合率は 0.0132 であった．

F 値にしても平均適合率にしても，表 1 や図 5 と比較すると，通常の教師付きの外れ値検出手法と同程度である．one-vs-rest 法を利用した場合でも，WSD システム単独では新語義の検出が困難であることがわかる．

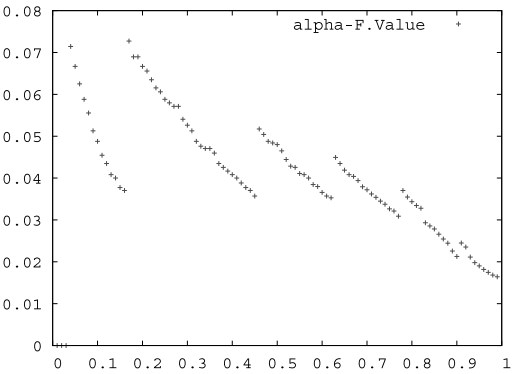

図 7 語義数を考慮した閾値と F 値

6.3 未出現語義を含めた評価

SemEval-2 日本語 WSD タスクでは，教師データ中には現れないが，テストデータには出現する語義が存在する．このような教師データ中の未出現語義は，新語義と見なすこともできる．このような用例は「あう」で 1 用例，「すすめる」で 1 用例，「出す」で 3 用例，「立つ」で 1 用例，「とる」で 3 用例，「ひとつ」で 1 用例，「見る」で 6 用例，「持つ」で 1 用例，「大きい」で 2 用例，「与える」で 1 用例の合計 20 用例存在する．これらも新語義の用例と見なした場合の検出結果を表 4 に示す．F 値の括弧内の数値は正解を新語義のみにした場合の正解数と F 値（表 1 の値）である．

また平均適合率の評価も行っておく．各手法の平均適合率の求め方は前述した方法で行う．結果を表 5 に示す．表 5 の「新語義のみ」の列は正解を新語義のみにした場合であり，「未出現語義を含む」の列は正解を新語義と未出現語義を合わせたものにした場合である．

表 4 未出現語義を含めた評価（F 値）

手法	抽出数	正解数	F 値
LOF	144	2 (2)	0.0222 (0.0250)
OCS	1228	10 (3)	0.0158 (0.0048)
OCS ∩ LOF	53	0 (0)	0.0000 (0.0000)
LOF-e	144	2 (2)	0.0222 (0.0250)
OCS-e	618	10 (3)	0.0306 (0.0095)
OCS ∩ LOF-e	50	0 (0)	0.0000 (0.0000)
NN	377	8 (6)	0.0387 (0.0305)
S-LOF	144	5 (4)	0.0556 (0.0500)
G-model	307	8 (5)	0.0466 (0.0310)
本手法	15	2 (2)	0.0784 (0.1290)

表 5 未出現語義を含めた評価（平均適合率）

手法	新語義のみ	未出現語義を含む
LOF	0.00507	0.00868
OCS	0.00386	0.01469
OCS ∩ LOF	0.00557	0.00883
LOF-e	0.00907	0.01613
OCS-e	0.00776	0.01619
OCS ∩ LOF-e	0.00869	0.01514
NN	0.01194	0.01782
S-LOF	0.01022	0.01578
G-model	0.01363	0.01896
本手法	0.03637	0.02930

表 6 未出現語義を含めた WSD の評価

手法	抽出数	正解数	F 値	平均適合率
識別の信頼度	39	2	0.0533 (0.0727)	0.01358 (0.00658)
one-vs-rest 法	33	1	0.0290 (0.0408)	0.01328 (0.00638)

F 値の評価でも平均適合率の評価でも本手法が最も高い値を出しており，本手法の効果は確認できる．ただし全体的な傾向として，未出現語義を正解に含めた場合の方が，F 値も平均適合率も若干高くなるが，本手法に関しては値が下がっている．S-LOF や G-model は未出現語義を正解に含めると，検出できる正解数は増えるが，共通して検出できる部分がなかったために，このような結果になった．この対策としては，後述するアンサンブル手法の導入により改良していきたい．

また，前節の WSD システムを用いた場合の評価を表 6 に示す．F 値と平均適合率の括弧内の数値は正解を「新語義のみ」にしたものである．

未出現語義を正解に含めた場合でも，前節同様，WSD システム単独では新語義の検出が困難であるといえる．

6.4 誤検出・未検出の原因

本手法の誤検出の原因について述べる．1 つは固有表現や熟語内の単語である．例えば以下のような表現が検出されている．

(a) 未来科学技術共同研究センターの中の研究施設
(b) 昔話の「千代ごこ出やっせ」のように
(c) 中小零細企業の取材は数多く手がかかる割りに

固有表現や熟語内の単語に通常の意味があるとは考えづらく，新語義の検出という観点では，このような表現を抽出しても完全に誤りとは言えない．本来，新語義の検出するためには，固有表現や熟語を予め抽出しておくことが必要だと考える．

また誤検出のその他の原因は多様であるが，全体として，対象単語の直前や直後に自立語が現れる複合語の用法や動詞の連体形の用法などが目立った．

(d) わが国が最も重要な貿易相手国の一つ
(e) 人間性を疑ってしまう人とは男女関係なく，
(f) 夏休み等に行って来た時の経験＝古き良き時代を，

複合語が専門性の高い用語である場合は意味のある検出とも見なせるが，ここでは複合語を単なる名詞連続で認識しているために，専門用語との区別は付けられない．新語義の検出に関しては，熟語や固有表現と同様，専門用語も通常の表現とは，区別した方がよいと考える．

本手法の未検出の原因としては，突き詰めれば，用例間の距離の測定方法に帰着される．ある新語義の用例と他の正常値の用例との距離がある程度，離れていたとしても，正常値の用例間の距離も同程度は離れているという状況である．これは動詞や形容詞における検出では顕著である．この問題に注目して距離学習を新語義発見に応用した研究も存在する (Sasaki and Shinnou 2012)．ただしこの問題は本質的に語義曖昧性解消の場合と同じであり，語義曖昧性解消の精度向上の試みが本研究に応用できる．

6.5 教師付き LOF とパラメータ k

オリジナルの LOF ではパラメータ k が存在し，この値が精度に大きく影響することが指摘されている．ここで提案した教師付き LOF では更に k の設定はシビアである．

教師付き LOF では，テストデータ y と最も近い点が教師データ x であった場合，x の密度が非常に高いために $LOF(y)$ の値も高くなり，一見，不都合に感じる．ただしテストデータ z の場合も，最も近い点が教師データ x であり，$d(x,y) < d(x,z)$ となっている場合は，$LOF(y) < LOF(z)$ となるために，y の外れ値の程度は z よりも下がる（図 8 参照）．

つまり極端に言えば，教師付き LOF は，最も近い点が教師データであり，しかもその点までの距離が大きい場合に外れ値の程度が大きくなる．これは外れ値の性質としては妥当である．現実的にはテストデータ y の k 近傍 $N_k(y)$ の中に教師データ x が入るかどうか，y から x までの距離 $d(x,y)$，$N_k(y)$ の中にテストデータがいくつ入るか及びそれらの位置関係が $LOF(y)$ の値に影響している．もしも $N_k(y)$ の中に教師データが入らない場合は，入る場合と比較して極端に $LOF(y)$ の値は小さいので，y が外れ値として検出されることはない（図 9 参照）．

「k 近傍内に教師データが入らない場合は外れ値ではない」という設定が妥当かどうかは不明である．当然，そうではない場合も想定することは可能だが，実験結果をみると本タスクにおいては上記設定が有効に機能していた．おそらく k 近傍内に教師データが入らない場合は，そのデータ近辺の密度が低いためだと考えられる．

ここで提案した教師付き LOF では，k 近傍内に教師データが入るかどうかで，外れ値かどう

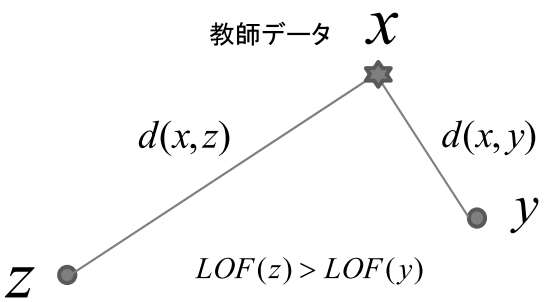

図 8 教師データとの位置関係による外れ値の度合い 1

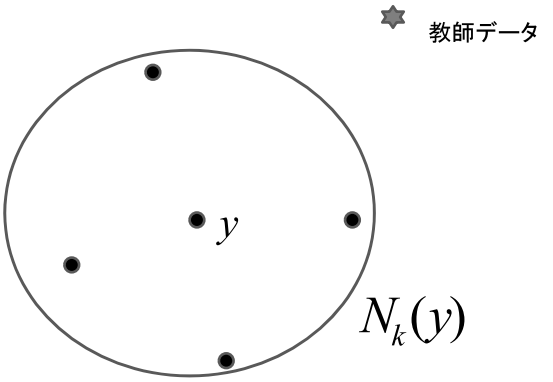

図 9 教師データとの位置関係による外れ値の度合い 2

かの最初の判定がされていると見なすこともできるので，パラメータ k の値は，通常の LOF よりも更に精度に影響を与えていると言える．

6.6 外れ値検出手法のアンサンブル

外れ値検出手法は数多く提案されており，本論文で利用した LOF についてもいくつかの改良手法が提案されている (Jin, Tung, Han, and Wang 2006; Papadimitriou, Kitagawa, Gibbons, and Faloutsos 2003)．これらの手法をどのようにして教師付きの枠組みへ拡張するかは不明であるが，これらを利用することで本手法の改善も可能である．

また，新たに外れ値検出手法を考案するのではなく，既存の手法を組み合わせる戦略も有効である．Lazavic は複数の外れ値検出の手法を適用して，それら出力結果を総合的に判断して最終的に外れ値候補を出力するという外れ値検出手法のアンサンブル (ensemble) を提案した (Lazarevic and Kumar 2005)．ここで提案した LOF と生成モデルの組み合わせも，外れ値検出手法のアンサンブルの一種と考えられる．ここでは単純に出力の積により最終の出力を決めたが，重みを付けて判断するなどの工夫も考えられる．あるいは他の外れ値検出の手法の組み合わせることも有効であろう．表 4 からもわかるとおり，LOF の出力と生成モデルの出力はかなり異なる．単純に出力の和を取ると，検出数が多くなりすぎて F 値の評価は下がってしまうが，第 1 段目の候補としては取り出せているので，そこからの選別に工夫することで改善が可能である．ここらが今後の課題である．

また，本論文では S-LOF と G-model のアンサンブルを提案したが，実験の結果をみると NN と G-model のアンサンブルや S-LOF と NN のアンサンブルも有望に見える．それらの実験結果を表 7 に示す[8]．

[8] G-model ∩ NN の平均適合率の測り方は G-model ∩ S-LOF（本手法）と同様である．つまり，G-model の出力の値を外れ値の度合いとし，本来の NN における出力のデータに対しては，G-model での出力の値に 100 を加えた後に，全体をソートした．

表 7 手法の組み合わせの評価

手法	抽出数	正解数	F 値	平均適合率
G-model ∩ NN	68	1	0.0238	0.01359
NN ∩ S-LOF	48	0	0.0000	0.01306
G-model ∩ S-LOF	15	2	0.1290	0.03637

表 7 が示すとおり，提案手法の S-LOF と G-model のアンサンブルが最も優れている．また組み合わせる手法によっては，個々の手法よりも精度が劣化することもありえるので，アンサンブルに用いる手法の選択も重要であることがわかる．

7 おわりに

本論文では対象単語の用例集合から，その単語の語義が新語義となっている用例を検出する手法を提案した．基本的に新語義の用例を用例集合中の外れ値と考え，外れ値検出の手法を利用する．ただし従来の外れ値検出では教師なしの枠組みであるが，ここではタスクの性質を考慮し，教師付きの枠組みで行った．

まず LOF を教師データを利用する形に改良した教師付き LOF を提案し，次に教師データを利用することで生成モデルを構築した．提案手法は上記 2 つの手法それぞれの出力の共通部分（積集合）を取るものである．これは 2 つの異なったタイプの外れ値検出の手法の積集合を取ることで誤検出を減らし，結果的に検出能力を高めることを狙いとしている．

タスクの一部として新語義識別を含む SemEval-2 の 日本語 WSD タスクのデータを利用して，LOF, One Class SVM, 最近傍法，教師付き LOF，生成モデルおよび提案手法による新語義の検出実験を行った．それぞれの手法の F 値と平均適合率を求めることで，提案手法の有効性を示した．また教師なしの手法 (LOF, OCS, OCS ∩ LOF)，単純に教師データを検出結果から除く手法 (LOF-e, OCS-e, OCS ∩ LOF-e) 及び教師付きの手法 (NN, S-LOF, G-model) の F 値と平均適合率を比較することで，新語義検出を目的とした外れ値検出では，教師データを積極的に利用することが精度向上に効果があることが確認できた．また WSD システムの識別の信頼度を利用した新語義を検出実験も行った．十分なパフォーマンスを示す WSD システムを用いても，WSD システム単独では新語義の検出が困難であることも示した．

提案手法は外れ値検出手法のアンサンブルの手法と位置づけられる．提案手法における出力結果のアンサンブルは，積集合をとるという単純なものであるため，この部分に工夫を入れることで更に検出能力が高まると予想している．出力結果の統合方法を工夫することが今後の課題である．

参考文献

Agirre, E. and Soroa, A. (2007). "Semeval-2007 task 02: Evaluating word sense induction and discrimination systems." In *SemEval-2007*.

赤穂昭太郎 (2008). カーネル多変量解析. 岩波書店.

Bordag, S. (2006). "Word sense induction: Triplet-based clustering and automatic evaluation." In *EACL-2006*, pp. 137–144.

Breuning, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). "LOF: Identifying Density-Based Local Outliers." In *ACM SIGMOD 2000*, pp. 93–104.

Denkowski, M. (2009). "Survey of Techniques for Unsupervised Word Sense Induction."

Erk, K. (2006). "Unknown word sense detection as outlier detection." In *NAACL-2006*, pp. 128–135.

Jin, W., Tung, A. K. H., Han, J., and Wang, W. (2006). "Ranking outliers using symmetric neighborhood relationship." In *The 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD '06)*, pp. 577–593.

九岡佑介，白井清昭，中村誠 (2008). 複数の特徴ベクトルのクラスタリングに基づく単語の意味の弁別. 第 14 回言語処理学会年次大会, pp. 572–575.

Lazarevic, A. and Kumar, V. (2005). "Feature bagging for outlier detection." In *The eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (KDD '05)*, pp. 157–166.

Okumura, M., Shirai, K., Komiya, K., and Yokono, H. (2010). "SemEval-2010 Task: Japanese WSD." In *The 5th International Workshop on Semantic Evaluation*, pp. 69–74.

Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C. (2003). "LOCI: Fast Outlier Detection Using the Local Correlation Integral." In *ICDE-2003*, pp. 315–326.

Sasaki, M. and Shinnou, H. (2012). "Detection of Peculiar Word Sense by Distance Metric Learning with Labeled Examples." In *LREC-2012*, pp. Session–P6.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). "Estimating the support of a high-dimensional distribution." *Neural Computation*, **13** (7), pp. 1443–1471.

Schütze, H. (1998). "Automatic word sense discrimination." *Computational Linguistics*, **24** (1), pp. 97–123.

Shinnou, H. and Sasaki, M. (2010). "Detection of Peculiar Examples using LOF and One Class SVM." In *LREC-2010*.

Shirai, K. and Nakamura, M. (2010). "JAIST: Clustering and Classification Based Approaches for

Japanese WSD." In *The 5th International Workshop on Semantic Evaluation*, pp. 379–382.

Sugiyama, K. and Okumura, M. (2009). "Semi-supervised Clustering for Word Instances and Its Effect on Word Sense Disambiguation." In *The 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009)*, pp. 266–279.

高村大也 (2010). 言語処理のための機械学習入門. コロナ社.

田中博貴，中村誠，白井清昭 (2009). 新語義発見のための用例クラスタと辞書定義文の対応付け. 第 15 回言語処理学会年次大会, pp. 590–593.

山西健司 (2009). データマイニングによる異常検知. 共立出版.

略歴

新納　浩幸：1985 年東京工業大学理学部情報科学科卒業．1987 年同大学大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，1993 年 4 月茨城大学工学部システム工学科助手．1997 年 10 月同学科講師，2001 年 4 月同学科助教授，現在，茨城大学工学部情報工学科准教授．博士（工学）．機械学習や統計的手法による自然言語処理の研究に従事．言語処理学会，情報処理学会，人工知能学会 各会員．

佐々木　稔：1996 年徳島大学工学部知能情報工学科卒業．2001 年同大学大学院博士後期課程修了．博士（工学）．2001 年 12 月茨城大学工学部情報工学科助手．現在，茨城大学工学部情報工学科講師．機械学習や統計的手法による情報検索，自然言語処理等に関する研究に従事．言語処理学会，情報処理学会各会員．

（2012 年 2 月 6 日　受付）
（2012 年 6 月 8 日　再受付）
（2012 年 7 月 18 日　再々受付）
（2012 年 9 月 20 日　採録）