科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　１２１０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００１６７

研究課題名（和文）外れ値検出手法を利用したコーパスからの新語義発見

研究課題名（英文）Detection of new word senses from a corpus by the outlier detection method

研究代表者
新納　浩幸（Shinnou, Hiroyuki）

茨城大学・工学部・准教授

研究者番号：１０２５０９８７

交付決定額（研究期間全体）：（直接経費）　　3,800,000円、（間接経費）　　1,140,000円

研究成果の概要（和文）：本研究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。ここでは、新語義の用例が用例集合中の外れ値になると考え、外れ値検出の手法を利用した。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、本タスクの場合、教師付きの設定の方が自然であり、その設定で行った。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た。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。

研究成果の概要（英文）：In this research, we proposed 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. Then using outlier detection methods, we detect the new word senses. Generally, outlier detection methods are considered to be unsupervised. However, supervised method is natural for our task. Therefore, our outlier detection method is classified under the supervised framework. We proposed 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 demonstrated the effectiveness of our method using SemEval-2 Japanese WSD task data.

研究分野：　情報学

科研費の分科・細目：　知能情報学

キーワード：　新語義　外れ値検出　LOF　生成モデル　SemEval

１．研究開始当初の背景

本研究では、対象単語の語義が既存の語義とは異なる意味（新語義）で使われている用例をコーパス（10 年分の新聞記事のような大量のテキスト文書）から自動発見する手法を開発する。

例えば、単語「ソース」を既存の辞書（岩波辞書）で調べると、(1) 西洋料理に調味料として使う汁。(2) 出所。源泉。の２つの語義が記載されている。一方、コーパスには『C 言語の文法をマスターしてもソースを解読するのは困難です。』という「ソース」の用例が存在する。この用例中の「ソース」の語義は「プログラムコード」であり、上記の既存辞書中の語義には存在せず、新語義と考えられる。「ソース」という対象単語が与えられたときに、このような新語義の用例を自動発見することが本研究のタスクである。

新語義を発見することは辞書を拡充、更新する際に有益である、また通常の語義識別問題を解決するには語義タグ付きの用例集を利用して機械学習手法を用いることが一般的であるが、その語義タグ付きの用例集を構築する際にも、新語義の用例を予め検出しておくことは有益である。また単語の使用法は時代と共に変化し、ある時点で新語義が現れる。このような言語の変化を調べることにも利用できる。また新語義として検出した用例は、実際は誤用であることも多く、誤り検出としても利用できる。

２．研究の目的

本研究の目的は、対象単語の語義が新語義であるような用例をコーパスから自動発見することである。そのためにデータマイニング分野の外れ値検出の手法を利用する。ただし通常の外れ値検出の手法は教師なし学習の枠組みであるが、ここではより現実的に少量の教師データが利用できるという設定で外れ値検出の手法を応用する。

本研究のアプローチは、データマイニング分野の外れ値検出の手法を利用して新語義の用例を発見するというものである。これは用例中の対象単語をベクトル（用例ベクトル）として表現した場合、新語義となる用例ベクトルは、既存語義の用例ベクトルに対する外れ値とみなせるという考えに基づいている。新語義発見に既存の外れ値検出の手法を利用する場合、何も手を加えず、そのまま利用することも可能であるが、新語義発見のタスクは従来の外れ値検出のタスクとは異なった特徴があり、従来手法をそのまま用いても良い結果は得られない。それは新語義発見のタスクでは外れ値の定義が明確である点である。通常、外れ値検出のタスクでは、外れ値の定義が曖昧である、つまりこれは外れ値ではない正常値の定義が曖昧であることを意味する。そしてほぼすべてのデータが正常値であるために、正常値と外れ値を識別するための教師データが存在しえない。そのため外れ値検出のタスクは教師なし学習の枠組みでしか捉えられてこなかった。しかし新語義発見のタスクでは、既存の辞書に存在しない語義を外れ値と定義できる。そしてこの定義に従えば、正常値のデータの一部に語義の情報を付与することができるし、そのような情報を利用できるとした問題設定の方が、現実の問題として自然である。つまり本研究は外れ値検出の手法を利用して、新語義発見を行うが、既存の外れ値検出の手法とは異なり、少量の教師付きデータ（語義タグ付きの用例）を利用する。

従来の外れ値検出の手法は多岐にわたるが、大きく分類するとデータの生成に確率モデルを用いるものと、用いないものに分けられる。確率モデルを用いた場合、データの生成確率が得られるので、その確率が低いデータを外れ値とする。このアプローチでは、いかに適切な確率モデルを導入できるかが鍵となる。確率モデルを用いない手法としては Local Outlier Factor (LOF) や One Class SVM が代表的である。LOF は密度ベースの手法であり、概略、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。One Class SVM は u-SVM を利用した外れ値検出の手法である。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、u-SVM を使って 2 つのクラスを分離する超平面を求める。その結果、-1 のクラス側に属するデータを外れ値とする。

本研究では基本的に確率モデルを用いないアプローチを取る。本研究の場合、少量の語義タグ付きの用例が利用できるという設定のために、確率モデルを用いるアプローチは、従来の語義曖昧性解消問題に対する帰納学習手法と等価となる。もちろんそのアプローチでも新語義の検出は可能かもしれないが、実際は難しいと考える。それは語義曖昧性解消問題に対する帰納学習手法では識別の困難な用例を新語義の用例と判定するが、識別の困難な用例が新語義である保障がないからである。

本研究では LOF と One Class SVM を少量の語義タグ付きの用例が利用できるという設定、つまり教師付きの枠組みで捉え直し、新語義発見のタスクに応用する。具体的には LOF も One Class SVM もタグの付いている用例に、ある種の重みをつけ、通常のタグのない用例とマージさせてから外れ値検出を行う。LOF の場合、重みをつけることで、タグの付いている用例の密度が高くなる。One Class SVM では識別の境界をタグの付いている用例から原点方向にずらすことになる。結果として、どちらの手法でも外れ値検出の精度が改善される。最終的には LOF と One Class SVM の両者の出力を組み合わせて、最終的な検出を行う。その際に

は確率モデルを用いた場合に得られる情報も利用する。

以上より、本研究では研究期間内に以下の4点の研究を行うことを目的とする。

・語義識別問題の従来手法を新語義発見に用いる際の問題点を明らかにする
・LOF におけるタグ付き用例の利用方法を提案する
・One Class SVM におけるタグ付き用例の利用方法を提案する
・確率モデル、LOF、One Class SVM の検出結果を総合して最終的な検出結果を求める

研究期間内で上記4点に関する研究結果を発表する。また本手法に対する実験データも必要に応じて作成してゆき、それも公開する。最終的には本研究により発見された新語義とその用例も公開する。

３．研究の方法

本研究目的を達成するために、平成 23 年度に、語義識別問題の従来手法を新語義発見に用いる際の問題点を明らかにする。そのために語義タグ付きの用例と帰納学習手法の Naive Bayes を用いて語義識別器を作成し、それを利用して新語義発見を行う。平成 24 年度に、LOF におけるタグ付き用例の利用方法及び One Class SVM におけるタグ付き用例の利用方法を提案する。具体的にはタグ付き用例に重みをつける方法を検討する。平成 25 年度に、確率モデル、LOF、One Class SVM の検出結果を総合して最終的な検出結果を求める手法を開発する。

まず平成 23 年度は、語義識別問題の従来手法を新語義発見に用いる際の問題点を明らかにする。具体的には語義タグ付きの用例と帰納学習手法の Naive Bayes を用いて語義識別器を作成し、それを利用して新語義発見を行う。この実験を通して、確率モデルによる新語義発見の問題点を指摘する。以下簡単に語義識別器の作成手順と、語義識別器を利用した新語義発見手法について述べる。語義のクラスを C = { c_1,c_2, ・・・, c_m } とする。用例 x が素性リスト (f_1,f_2,・・・,f_n) により表現されているとする。語義識別は確率 P(c|x) を推定することで実現できる。実際に x の語義 c_x は arg max P(c|x) により得られる。またベイズの定理より c_x = arg max P(c)P(x|c) となる。ここで事前確率 P(c) は教師データの語義の頻度の割合から推定する。P(x|c) に関しては Naive Bayes では P(x|c) = Π P(f_i|c) を仮定する。これによって x の語義 c_x が得られる。次にテストデータ y に対して P(c_y)P(y|c_y) を求め、この集合が正規分布であると仮定する。P(c_y)P(y|c_y) の値が小さい場合、識別されたクラスの信頼度が低いことを意味する。ここで極端に P(c_y)P(y|c_y) の値が小さいものが外れ値、つまり新語義と判定する。この際の閾値は正規分布の左片側 1% 分位点である -2.326 を用いる。

上記の手法により新語義発見の実験を行い、その問題点を明らかにする。実験で使うデータは SemEval-2 の日本語辞書タスクのデータを用いる。このタスクでは 50 単語の各対象単語に対して、教師データとテストデータが存在する。しかもテストデータには新語義も含まれているので、本研究の実験に利用できる。

平成 24 年度は、LOF におけるタグ付き用例の利用方法及び One Class SVM におけるタグ付き用例の利用方法を提案する。平成 23 年度の研究結果から、確率モデルで生じる問題を LOF や One Class SVM により解決する。どちらの手法を用いる場合でも、解決の鍵となるのは教師データに対応する語義タグ付き用例の利用法である。語義タグ付き用例には、既存語義のタグが付いているので、当然新語義の用例として検出されてはいけない。LOF や One Class SVM は本来教師なし学習の一種なので、教師データを必要としない。このため単純に語義タグ付き用例を検出の対象から外すだけでは、教師データを全く利用しないことになる。そこで外れ値検出の精度を向上させるように教師データを利用する方法を検討する。

本研究の申請時に考案したアイデアを簡単に述べる。基本的にこの方向から研究を始めた。

LOF では教師データとテストデータ（検出対象のデータ）を合わせたデータからの検出を行う。ただし、このままだと教師データからも外れ値を検出する可能性がある。しかも教師データは外れ値ではないという情報を全く利用していない。そこで教師データの密度を高くする設定を行う。LOF では密度を測る基本アイデアとしてデータ x に対して kdist(x) という値をを定義している。直感的に kdist(x) は x から k 番目に近いデータまでの距離である。LOF では kdist(x) によってデータ x の密度に相当するものを測る。本研究では教師データを k 倍して検出対象データに加えることを検討する。ここで k は kdist(x) の k である。これによって教師データからは新語義を検出することがなくなり、しかも教師データの近傍のデータも同時に密度が高まるので、それらのデータも検出されなくなる。これによって外れ値検出の精度が高まる。

４．研究成果

本研究の成果は3件の雑誌論文、19 件の学会発表にまとめられる。その中でも、以下の論文は、本研究成果全体をまとめたものである。この論文の内容を概説することで、ここ

での研究成果の報告とする。

新納浩幸、佐々木稔、『外れ値検出手法を利用した新語義の検出』、自然言語処理、Vol.19、No.4、pp.303-327 (2012)

本研究では、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した。

まず教師付き外れ値検出について説明する。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。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おえない。

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。

本研究ではこれらの特徴を利用して外れ値検出を行った。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った。

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、本研究は k=5 とした。

次に教師データを利用した生成モデルの構築について述べる。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 としたとき、全確率の公式から $P(x)=\Sigma P(z_i) P(x|z_i)$ が成立する。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、$\Sigma n_i = N$ であり、$P(z_i)=n_i/N$ と推定できる。問題は $P(x|z_i)$ の推定である。x は素性リスト $x=\{f_1,f_2,\cdots,f_m\}$ で表現されている。ここでは Naive Bayes で使われる素性間の独立性を仮定して、$P(x|z_i)=\Pi P(f_j | z_i)$ 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、$P(f_j|z_i)= n(z_i,f_j)/n_i$ と推定できる。これまで得られた式を MAP 推定でスムージングを行い、$P(z_i)=(n_i+1)/(N+K)$ と $P(f_j|z_i)=(n(z_i,f_j) + 1)/(n_i + 2)$ と補正する。以上より P(x) の値が求まる。外れ値の度合いは -logP(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 -\logP(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 -logP(x) を計算し、それらの値に対する平均 m と分散 s2 を求める。-logP(x) の分布を正規分布と考え、正規化した値に対して閾値を設けることにした。正規化した値が閾値以上の x を外れ値とした。

最後に教師付き LOF と生成モデルの積集合をとり、最終的な出力とする。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。

実験は SemEval-2 の日本語辞書タスクのデータを用いて行った。このタスクは通常の語義曖昧性解消のタスクだが新語義がテストデータ中に存在するという特徴があり、本研究の提案手法の評価に使える。ここではこのデータを用いて、新語義検出の F 値、平均的適合率による評価を行った。いくつかの従来手法を実装し、提案手法と比較することで、提案手法の有効性を示した。

また提案手法に対して、教師データを k+1 倍する効果も確認した。また通常の語義曖昧性解消のシステムを利用するだけでは新語

義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手法の一種であるが、アンサンブルする方法が簡易である。アンサンブルする方法を工夫して検出精度を高めることが今後の課題である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① 新納浩幸、佐々木稔、『共変量シフトの問題としての語義曖昧性解消の領域適応』、自然言語処理、Vol.21、 No.1、 pp.61-79 (2014)．査読有
② 新納浩幸、佐々木稔、『k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応』、自然言語処理、Vol.20、No.5、pp.707-726 (2013) 査読有
③ 新納浩幸、佐々木稔、『外れ値検出手法を利用した新語義の検出』、自然言語処理、Vol.19、No.4、pp.303-327 (2012) 査読有

〔学会発表〕（計 19 件）

① 新納浩幸、國井慎也、佐々木稔、『語義曖昧性解消を対象とした領域固有のシソーラスの構築』、第 5 回日本語学ワークショップ、pp.199-206 (2014).(立川)
② 菊池裕紀、新納浩幸、『uLSIF による重み付き学習を利用した語義曖昧性解消の領域適応』、第 5 回日本語学ワークショップ、pp.63-70 (2014)．(立川)
③ 小野寺善行、新納浩幸、『領域間距離を利用した能動学習による語義曖昧性解消の領域適応』、第 5 回日本語学ワークショップ、pp.57-62 (2014)．(立川)
④ 吉田拓夢、新納浩幸、『外れ値検出手法を利用した Misleading データの検出』、第 5 回日本語学ワークショップ、pp.49-56 (2014)．(立川)
⑤ Shinya Kunii and Hiroyuki Shinnou、『Combined Use of Topic Models on Unsupervised Domain Adaptation for Word Sense Disambiguation』、 PACLIC-27、pp.415-422 (2013)． (台湾)
⑥ 吉田拓夢、新納浩幸、『語義曖昧性解消の領域適応における Misleading データの存在と検出』、第 4 回日本語学ワークショップ、pp.317-324(2013)．(立川)
⑦ 小野寺善行、新納浩幸、『クラスタリングを利用した能動学習による語義曖昧性解消の領域適応』、第 4 回日本語学ワークショップ、pp.309-316 (2013)．(立川)
⑧ Horonori Kikuchi and Hiroyuki Shinnou、『Domain Adaptation for Word Sense Disambiguation under the Problem of Covariate Shift』、NL-212-4 (2013)．(函館)
⑨ Shinya Kunii and Hiroyuki Shinnou、『Combined Use of Topic Models on Unsupervised Domain Adaptation for Word Sense Disambiguation』、NL-212-3 (2013)．(函館)
⑩ 新納浩幸、 佐々木稔、『k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応』、情報処理学会自然言語処理研究会、 NL-211-13 (2013).(品川)
⑪ 小幡智裕、佐々木稔、新納浩幸、『サポートベクターマシンに基づく Hit Miss Network を用いたインスタンス選択』，言語処理学会第 19 回年次大会、P6-11 (2013)．(名古屋)
⑫ 國井慎也、新納浩幸、佐々木稔、『ミドルソフトタグのトピック素性を利用した語義曖昧性解消』、言語処理学会第 19 回年次大会、P3-9 (2013)．(名古屋)
⑬ Minoru Sasaki and Hiroyuki Shinnou、『Word Sense Disambiguation Based on Distance Metric Learning from Training Documents』、The Sixth International Conference on Advances in Semantic Processing (2012)．(スペイン)
⑭ Minoru Sasaki and Hiroyuki Shinnou、『Detection of Peculiar Word Sense by Distance Metric Learning with Labeled Examples』 LREC-2012 (2012)．(トルコ)
⑮ 新納浩幸、佐々木稔、『外れ値検出手法を利用した新語義の検出』、言語処理学会第 18 回年次大会、 D5-5 (2012)．(広島)
⑯ 真下飛瑠、新納浩幸、佐々木稔、『逆トピックワードを利用した外れ値文書検出』、言語処理学会第 18 回年次大会、 P3-35 (2012)．(広島)
⑰ 西野太樹、新納浩幸、佐々木稔、『トピックモデルを用いた語義曖昧性解消』、言語処理学会第 18 回年次大会、P3-8 (2012)．(広島)
⑱ 佐々木稔、新納浩幸、『商品タイトルから商品名を自動抽出するための効率的な教師データ作成手法』、言語処理学会第 18 回年次大会、E4-1 (2012)．(広島)
⑲ 新納浩幸、全太俊、佐々木稔、『人名構成文字確率を用いた文字ベース CRF による中国語人名検出』、言語処理学会第 18 回年次大会、P2-31 (2012)．(広島)

６．研究組織
(1)研究代表者
新納　浩幸（SHINNOU HIROYUKI）
茨城大学・工学部・准教授
研究者番号：10250987

(2)研究分担者
無し

(3)連携研究者
無し