科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月 １７ 日現在

機関番号：１２１０１
研究種目：若手研究(B)
研究期間：2012～2013
課題番号：２４７００１３１

研究課題名（和文）単語の語義別コロケーション抽出とその語義識別、新語義発見への適用に関する研究

研究課題名（英文）Research on method of extracting collocation for each word sense and its application for word sense disambiguation

研究代表者
佐々木　稔（Sasaki, Minoru）

茨城大学・工学部・講師

研究者番号：６０３４４８３４

交付決定額（研究期間全体）：（直接経費）　　1,600,000円、（間接経費）　　480,000円

研究成果の概要（和文）：平成24年度は、「訓練データからの語義別コロケーション抽出と、それを考慮した語義識別モデル構築法の検討」についての研究を行った。語義別コロケーションの検出や語義の識別手法などについて、学術誌論文1編と国際会議発表2件、国内発表2件の成果を発表した。
平成25年度は、「語義別コロケーションを考慮した共起ベクトルの構築手法の検討」について研究を行った。コロケーションなどの複数語の共起を既存手法で得られる単語共起ベクトルの空間に組み込む手法などについて、学術誌論文2編と国際会議1件、国内発表2件の成果を発表した。

研究成果の概要（英文）：In 2012, I study a method of extracting collocation for each word sense from the training data and a method of identifying the senses of words in text using collocation dictionary. As a result, I publish one journal article, two international conference papers and two domestic conference papers about the method of collocation extraction for each word sense and word sense disambiguation.
In 2013, I study a method of method of extracting collocation for each word sense and its application for word sense disambiguation. I publish two journal papers, one international conference paper and two domestic conference papers about a word sense disambiguation method based on the global co-occurrence information.

研究分野：複合領域

科研費の分科・細目：情報学、知能情報学

キーワード：自然言語処理

1．研究開始当初の背景

ある単語が含まれる用例文集合に対して、語義別に用例文を分類することは本格的な意味解析を行う上で非常に有用なデータセットの構築へとつながる。例えば、語義別に分類された用例文集合が存在すれば、語義ごとに周辺の共起語を分析することで語義識別モデルを作成し、単語の意味を特定するための分類器を作ることができる。また、動詞についての格フレームを容易に自動構築することや語義ごとに項目分けをしたシソーラスを容易に構築することなどに有効であると考えられる。

語義別に用例文を分類する際に、用例文集合の中には辞書の分類項目には存在しない語義による用例や誤った用法による用例といった、特異な用例文が数多く存在する。そのような特異用例を検出することで、語義識別問題に向けた訓練データの作成や既存の辞書における分類項目の改定などに役立てることができる。しかし、上記のように特異用例には様々な種類が存在しているため、特異用例かどうかを主観的な観点で判断することになり、従来の識別問題として解くことが非常に困難な状況にある。

これまで、申請者は文部科学省科学研究費特定領域研究「日本語コーパス」の研究項目「代表性のあるコーパスを利用した日本語意味解析」の研究分担者として参加し、単語の新語義、新用法の自動発見手法の開発に携わってきた。この研究テーマでの議論において、特異用例の中でも評価することが容易な特異用例として、辞書の分類項目には存在しない語義を新語義と定義し、使用された用例文の抽出を行ってきた。その手法として、用例文の周辺単語から得られる特徴ベクトルの分布密度から外れ値を検出する Local Outlier Factor(LOF) と識別手法を応用して外れ値を検出する One-Class Support Vector Machine (OCSVM) の組合せを訓練データに適用し、新語義の抽出を行う手法を提案し、LOF や OCSVM を個別に使う場合と比較して新語義抽出精度を向上させることができた。しかし、この手法は既知の語義情報を利用していないため、語義情報を利用した外れ値検出手法を開発することが課題となる。また、識別問題としての識別精度を上げるアプローチについても検証する必要がある。

同じく上記プロジェクトの共同研究者である北陸先端科学技術大学院大学の白井清昭准教授のグループも、我々とは異なるアプローチで用例文から特徴ベクトルを構成し、新語義をひとつのクラスとする分類手法を提案している。この手法では、語義識別精度は高いが新語義の判定精度が低い結果となり、同様に識別モデルの構築に課題が残っている。

また、国際ワークショップ Semeval2010 の日本語語義識別タスクにおいて提出されたシステムについて、新語義の正解率が最も高い要因を調査した。その中で、訓練データ内に新語義のタグが付与された単語が存在する時のみ新語義ラベルを考慮した識別モデルを作ることで、新語義発見の性能改善が見られた。訓練データ内で新語義のタグが付与されているのはキーワード「可能」に対して 50 用例文中 14 件のみ(すべて「可能性」の一部)であった。これらの訓練データを利用して学習を行うだけで高い正解率が得られる理由は、新語義は特定のフレーズとして用いられることが多いからであると考えられる。そのため、頻出するフレーズが語義形成の大きな要因になると仮定することができる。

上記のように、申請者やその他の研究者によるこれまでの研究成果において、新語義を発見するためには既存語義を現状よりもさらに高い精度で識別できるシステムの構築が求められている。また、新語義は決まったフレーズとして用いられることが多く、決まったフレーズを効果的に利用することが新語義抽出精度の向上に有効な情報であると考えられる。

2．研究の目的

上記に示した状況を踏まえ、広い概念を持つ単語でも単語の組合せで概念が限定し、意味分類に強く影響することに着目し、このような単語の組合せを文書集合から抽出し、知識として利用することで語義識別の性能向上と新語義発見の精度改善を図る。そのために、

(1) 「訓練データからの語義別コロケーション集合の抽出手法の検討」、
(2) 「語義別コロケーション集合を考慮した語義識別モデル構築法の検討」、
(3) 「語義識別モデルからの新語義発見法の検討」

の 3 段階に分けて問題の検討を行う。

特に、段階 1．では既存システムでは考慮されていない、ラベル付きデータからコロケーションを特定する問題について議論が必要となる。また、段階 2．では語義別コロケーションを利用した効果的な語義識別モデル構築問題について議論を行う。段階 3．では既存語義の識別能力を保ちつつ、新語義の発見精度を向上するための問題について議論を行う。

3．研究の方法

(1) 訓練データからの語義別コロケーション抽出と、それを考慮した語義識別モデル構築方法についての検討を行った。

語義別コロケーション抽出では、新語義も含めたあまり出現しない語義の用例抽出を行う必要がある。新語義として使用した用例を検出するためのアプローチのひとつとして、データマイニング分野で使われる外れ値検出手法の適用が挙げられる。これは、用例集合の中で新語義の用例が特殊な単語の使い方をしていると考え、用例集合の中から外れ値を抽出することでそれが新語義の用例であると判定する手法である。ただし、ここで使われる外れ値検出手法はラベル情報を利用しない教師なし検出手法で、データの密度を利用して検出が行われる、そのため、新語義の検出で利用可能な教師データの語義情報を利用できず、語義情報も考慮した用例間の関連性を捉えることができない問題があった。

そこで、データの一部にラベルが割り当てられた集合に対して、ラベル情報も考慮した外れ値の検出手法を提案し、用例集合から新語義として使用した用例候補の検出を行う。この手法はラベル付きデータに対し、距離学習手法のひとつである Large Margin Nearest Neighbor (LMNN) を利用して同じラベルを持つデータは近くに集め、異なるラベルを持つデータは遠くに移動することにより、ラベル情報を考慮したデータの分布を求める。この距離学習を行ったデータ集合に、外れ値の指標である Local Outlier Factor(LOF) を利用することで外れ値候補の抽出を行った。

提案手法の有効性を評価するために、人工的に生成したデータによる外れ値検出を行う実験と Semeval-2010 日本語 WSD タスクのデータによる新語義用例検出を行う実験を行った結果、提案手法は外れ値の検出件数、および、F 値で LOF、One-Class SVM を上回る検出結果となり、密度に基づく新語義検出において、教師データの利用が有効であることが分かった。また、多くの用例について学習後に LOF 値の順位が上がり、距離学習による密度変化が新語義検出に有効であることが分かった。

(2) 用例文の特徴と語義ラベルを利用した用例文間の距離学習手法についての検討を行った。

一般的にベクトル空間モデルを基本とした語義識別は、ある単語について同じ語義を持つ場合にはその単語の周辺において共起する単語の出現傾向が類似していると言われる。また、異なる語義で単語を使う場合には、一方の語義と比較して異なる単語が出現する傾向にある。距離学習手法は同じ語義を持つ特徴ベクトルの点集合は近い場所に集め、異なる語義を持つ点は遠い場所に離すことで、より語義識別しやすい特徴ベクトルを獲得するものである。

正解語義が割り振られた用例文集合に対し、用例文間の類似性を測定するために、座標軸変換による距離学習手法である Local Fisher Discriminant Analysis (LFDA)、Semi-Supervised Local Fisher Discriminant Analysis (SELF)と、データの移動による距離学習手法である Neighborhood Component Analysis (NCA) と Large Margin Nearest Neighbor (LMNN) を利用する場合について語義識別実験を行った。

その結果、LMNN を利用した語義識別手法を利用することで、従来よく利用される SVM よりも高い精度で識別することが可能であることを示した。また、LMNN を利用した場合は、3 つ程度の特定の学習データのみで語義を識別する傾向や 3 つ以上の語義を持つ場合の各語義間の関係を調べる上で有効な手段であることが分かった。

(3) 語義別コロケーションを考慮した共起ベクトルの構築手法について検討を行った。

学習データを持っている場合、語義曖昧性解消は分類問題として定式化され、教師あり学習手法を用いて解くことが多い。この教師あり学習手法を適用するためには、学習データの特徴として単語集合を抽出し、頻度などの重みにより数量化した共起ベクトルで特徴表現が行われ

る。しかし、このような特徴表現は局所的な周辺文脈情報が使われているが、文書集合全体に含まれるグローバルな共起情報をベクトルに組み込むことができず、この情報は使われていなかった。そのため、広い範囲の文書集合で使われるコロケーション情報を、前後の単語共起と組み合わせてベクトルで表現するのは難しい問題であった。

そこで、グローバルな共起情報を単語共起ベクトルに効果的に組み込むために、一般的な行列分解手法である Non-Negative Matrix Factorization (NMF) を利用することを提案し、語義曖昧性解消を行った。従来研究では、対象単語が含まれる文の中での共起情報を組み込んでいたが、スパースな行列であるために NMF が望ましい解に収束しない場合があった。それを解決するために、文書集合全体で出現する共起関係を NMF を利用して組み込んだ。これにより、NMF が収束しない問題が解決し、より効果的な特徴を導出することができる。

この手法を利用して語義曖昧性解消実験を行った結果、対象単語前後の共起を使う従来手法よりも高い精度で曖昧性を解消することができた。また、文書集合全体から抽出した共起情報を利用することで、語義曖昧性解消に安定した効果が得られることがわかった。

４．研究成果

(1) 研究の主な成果

① 「訓練データからの語義別コロケーション抽出と、それを考慮した語義識別モデル構築法の検討」では、新語義も含めたあまり出現しない語義の用例抽出を行う必要があるため、用例文集合から対象単語が特異な使用をしている用例を検索する手法の開発を行った。この手法は、国際会議 LREC2012 において発表し、低頻度語義を持つ用例の特徴を分析し、それを含む用例を効率的に抽出することを実証した。

② 「用例文の特徴と語義ラベルを利用した用例文間の距離学習手法についての検討」では、訓練データからの語義別コロケーションを抽出するためには、正解語義が割り振られた用例文集合に対し、用例文間の類似性を測定することが重要な課題となる。そのため、用例間類似度を語義ラベルに応じて学習する手法の開発を行った。この研究成果は、国際会議 SEMAPRO2012 において発表し、開発した用例間類似度を利用することで、従来の類似度尺度よりも高い語義識別精度が得られたことを実証した。

③ 「語義別コロケーションを考慮した共起ベクトルの構築手法についての検討」では、コロケーションなどの複数語の共起を既存手法で得られる単語共起ベクトルの空間に組み込む手法の開発を行った。「Semeval2010 日本語語義曖昧性タスク」で語義識別実験を行った結果、この手法を利用して複数語からなるコロケーション情報が語義識別に対して効果的であることを実証した。この手法は国際会議 NLP-2014 において発表し、この手法の妥当性を示した。

(2) 得られた成果の国内外における位置づけとインパクト

上記の成果は、これまであまり分析が行われていなかった低頻度語義用例の特徴を明らかにし、それを含む用例の抽出を効果的に行う手法構築することである。従来技術では 70%～80%の精度で頭打ちの状態であるが、そこで扱われる特徴以外に有効な特徴を見つける姿勢は、国内外の研究事例と比較して特色を有するのではないかと考えられる。また、実世界において使われる語義を網羅する目的において、語義別コロケーションを抽出するなどの今回の研究成果はある程度のインパクトがあると思われる。これらの知見は辞書編集を行う上でも意味の変化と転用の仕組みなどの分析や用例文の作成にも役立つと考えられる。

(3) 今後の展望

今後は得られた語義別コロケーションを利用して、既存の語義タグ付きコーパスの改良や修正を行うことが挙げられる。素性として扱うのは単語だけではなく、コロケーションを素性として決まった意味のまとまりをひとつの素性とするようにコーパスを修正することが可能となるかと思われる。

また、動詞についての格フレームを容易に自動構築することや語義ごとに項目分けをしたシソーラスを新規に構築する場合においても役に立つと考えられる。決まったフレーズを効果的に利用することが自然言語処理技術の向上に有効な情報であるため、幅広くコロケーション情報の利用が広がることを期待する。

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５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、２１巻、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、２０巻、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、１９巻、303-327、2013、査読有
④ Minoru Sasaki、Hiroyuki Shinnou、『Word Sense Disambiguation Based on Distance Metric Learning from Training Documents』、The Sixth International Conference on Advances in Semantic Processing、６巻、54-58、2012（イタリア）、査読有
⑤ Minoru Sasaki、Hiroyuki Shinnou、『Detection of Peculiar Word Sense by Distance Metric Learning with Labeled Examples』、The Eighth International Conference on Language Resources、８巻、601-604、2012（トルコ）、査読有

〔学会発表〕（計５件）

① 小幡智裕、佐々木稔、『インスタンス選択による文書データの効率的な分類モデル構築手法』、言語処理学会第２０回年次大会、2014.3.20、北海道大学、査読無
② 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回コーパス日本語学ワークショップ、2014.3.7、国立国語研究所、査読無
③ Minoru Sasaki、『Latent Semantic Word Sense Disambiguation Using Global Co-occurrence Information』、Third International Conference on Natural Language Processing、2014.2.21、Pullman Sydney（オーストラリア）、査読有
④ 小幡智裕、佐々木稔、『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、2013.3.15、名古屋大学、査読無
⑤ 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、2013.3.15、名古屋大学、査読無

〔その他〕
ホームページアドレス
http://sas.cis.ibaraki.ac.jp/

茨城大学研究者総覧
http://info.ibaraki.ac.jp/Profiles/5/0000495/profile.html

６．研究組織
(1)研究代表者
佐々木　稔（SASAKI　MINORU）
茨城大学・工学部・講師
研究者番号：60344834

(2)研究分担者
無し

(3)連携研究者
無し

(4)研究協力者
新納　浩幸（SHINNOU　HIROYUKI）
茨城大学・工学部・准教授
研究者番号：10250987