様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月　１日現在

機関番号：１２６０８
研究種目：特定領域研究
研究期間：２００６～２０１０
課題番号：１８０６１００３
研究課題名（和文）　代表性のあるコーパスを利用した日本語意味解析

研究課題名（英文）　Japanese semantic analysis using balanced corpus of contemporary Written Japanese

研究代表者
奥村　学　（OKUMURA MANABU）
東京工業大学・精密工学研究所・教授
研究者番号：６０２１４０７９

研究成果の概要（和文）：
語義タグ付コーパスの構築では,データ班から公開されているコアデータに対して，岩波国語辞典中の語義の区分に基づき，人手で語義を付与する作業を行った．BCCWJ を用いた新しい語義曖昧性解消タスクでは，語義曖昧性解消に関する評価型ワークショップであるSemeval-2（http://semeval2.fbk.eu/Semeval2.html）にBCCWJ を用いた語義曖昧性解消の評価型タスクを提案し，採択された．代表性のあるコーパスを用いた語義曖昧性解消では，ソースデータとターゲットデータの組み合わせごとに効果的な領域適応手法を自動的に選択する手法の開発を行っている．半教師ありクラスタリング手法の開発と，多義性解消への適用では，クラスタリング時に，教師情報を部分的に利用する，半教師ありクラスタリング手法を開発している．

研究成果の概要（英文）：
1) We constructed a corpus with word-sense annotation, based on the balanced contemporary corpus of written Japanese.
2) We organized the SemEval-2 Japanese Word Sense Disambiguation (WSD) task by using the corpus that we constructed in 1). Nine systems from four organizations participated in the task.
3) We showed that when domain adaptation for WSD (word sense disambiguation) was performed, the most effective domain adaptation method varies according to the properties of the source data and target data. We also presented the way to select the most effective method for domain adaptation depending on these properties using decision tree learning. The average accuracy of WSD showed significant improvement when the domain adaptation method which is selected automatically was used respectively, compared to when the original methods were used collectively.
4) We proposed a supervised word sense disambiguation (WSD) system that uses features obtained from clustering results of word instances. Our approach is novel in that we employ semi-supervised clustering that controls the fluctuation of the centroid of a cluster, and we select seed instances by considering the frequency distribution of word senses and exclude outliers when we introduce “must-link” constraints between seed instances. In addition, we improved the supervised WSD accuracy by using features computed from word instances in clusters generated by the semi-supervised clustering.
5) We proposed a method of detecting new word senses in a corpus. It consists of two procedures: (A) clusters of word instances are constructed so that the instances of the

same sense are merged, (B) then similarity between a cluster and a sense in a dictionary is measured in order to determine senses of instances in each cluster.

6) We proposed the method to detect peculiar examples of the target word from a corpus. Our method is to combine the density based method, Local Outlier Factor (LOF), and One Class SVM, which are representative outlier detection methods in the data mining domain. Our method improved precision and recall of LOF and One Class SVM. And we show that our method can detect new meanings by using the noun 'midori (green)'.

7) We presented a co-clustering-based verb synonym extraction approach that increases the number of extracted meanings of polysemous verbs from a large text corpus. Our proposed approach can extract the different meanings of polysemous verbs by recursively eliminating the extracted clusters from the initial data set. The experimental results of verb synonym extraction show that the proposed approach increases the correct verb clusters by about 50% with a 0.9% increase in precision and a 1.5% increase in recall over the previous approach.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	11,100,000	0	11,100,000
２００７年度	18,400,000	0	18,400,000
２００８年度	18,400,000	0	18,400,000
２００９年度	18,400,000	0	18,400,000
２０１０年度	18,400,000	0	18,400,000
総　計	84,700,000	0	84,700,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：語義タグ付コーパス,単語の新語義発見,機械学習,語彙概念構造,クラスタリング

１．研究開始当初の背景

日本語を対象にした言語処理研究では，形態素解析，構文解析について研究が進み，高精度なツールの開発も行われてきており，それらのツールが日本語学，日本語教育など他の研究分野でも広く利用されるようになってきている．その一方で，意味解析については依然研究が遅れており，一般に利用可能なツールの開発レベルにまで解析精度が到達していない．また，代表性のあるコーパスを用いた言語処理研究は，これまでそのようなコーパスが存在しなかったため，日本語に関してはまったく行われてこなかったと言って良い．

２．研究の目的

そこで本研究課題では，研究項目 A で構築する代表性のあるコーパスを用いた実証研究を行う．具体的には，以下の 3 つを柱とした日本語意味解析手法の開発を行う．

1）機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築

2）単語の新語義，新用法の自動発見手法の開発

3）語彙概念構造に基づく動詞の意味構造の自動抽出手法の開発と，それを用いた動詞の述語項構造辞書の自動構築手法の開発

3つの柱のうち，2）単語の新語義，新用法の自動発見手法の開発については新たな研究計画を平成21 年度より追加した．新たに追加する研究はコーパス中の特異な用例を検出する

手法の開発である．単語の特異な用例は，その単語の使われ方を調査する上で有用である．また特異な用例を検出・排除することで，用例集を精度良く分析することが可能となる．またコーパス内の特異な用例の有無を調べることで，そのコーパスの一般性や特殊性も考察できる．2）では当初，コーパス中の単語の用例集合をクラスタリングし，同じ意味を持つクラスタを作成した上で新語義を発見する手法を構想していた．しかし，この手法では，一定量同じ意味の用例が出現するまでクラスタが構成できず，したがって，新語義を発見できないという問題点があった．そのため，上述した特異な用例検出手法により，ごく少数の特異な用例しか出現していない時点でも新語義を発見できる手法を開発することで，2）で当初構想していた手法を補完し，新語義発見手法の完成度を増すことを狙っている．

３．研究の方法

本研究課題では，研究項目 A で構築する代表性のあるコーパスを用いた実証研究として，以下の 3 つの日本語意味解析手法の開発を行う．

1） 機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築

タグ付コーパスから学習した多義性解消システムによりタグ付コーパス作成コストの軽減を図るとともに，作成されたコーパスを用いて bootstrap 的に多義性解消システムの性能向上を図る．

2） 単語の新語義，新用法の自動発見手法の開発

時を経るにしたがって単語の意味は変化し，新しい意味が生まれることが知られている．今回構築されるような，時間幅を伴うコーパスで顕著に見られるこの言語現象を自動的に発見する手法を開発する．1)で開発する多義性解消手法で特定できない語義は新語義と考えられるため，2)は1)のシステムの自然な拡張と言える．

3） 語彙概念構造に基づく動詞の意味構造の自動抽出手法の開発と，それを用いた動詞の述語項構造辞書の自動構築手法の開発

語彙概念構造は動詞の振る舞いに関する分析から動詞の意味をそれが取る名詞同士の意味関係で記述する言語学に基づく意味構造である．文の意味構造は，1)で特定される単語の語義と 3)で抽出される意味構造の統合により得ることができる．

４．研究成果

1） 機械学習手法に基づく多義性解消手法の開発と，それを用いた代表性のある語義タグ付コーパスの半自動構築

東京工業大学の研究グループでは以下の4つを柱に研究を進めてきた．

_ 語義タグ付コーパスの構築，

_ BCCWJ を用いた新しい語義曖昧性解消タスク，

_ 半教師ありクラスタリング手法の開発と，多義性解消への適用，

_ 代表性のあるコーパスを用いた語義曖昧性解消．

語義タグ付コーパスの構築では，データ班から公開されているコアデータに対して，岩波国語辞典中の語義の区分に基づき，人手で語義を付与する作業を行った．語義付与対象単語は，岩波国語辞典中に見出し語がある単語で，かつ，複数の語義を持ち，品詞が，名詞，動詞，形容詞，副詞であるものとした．過去のタグ付コーパス構築例にならい，タグ付けの際，辞典中に該当の語義が見当たらない場合「該当なし」という判断を許し，また，

最下層の語義のどれかでは判断できない場合，より上位のラベルを付与することを許している．

BCCWJ を用いた新しい語義曖昧性解消タスクでは，語義曖昧性解消に関する評価型ワークショップである Semeval-2 (http://semeval2.fbk.eu/Semeval2.html) にBCCWJ を用いた語義曖昧性解消の評価型タスクを提案し，採択された．国内外の合計10グループが参加表明をしていたが，2010 年の3 月から4 月にかけて行われたformal run では最終的に，領域内の2 グループと領域外の2 グループ(海外からの1 グループを含む)が結果を提出した．このタスクの参加者も含め，現時点では合計で領域内5 グループ，領域外の国内4 グループ，領域外の海外4 グループが，構築した語義タグ付コーパスを利用していることになる．

半教師ありクラスタリング手法の開発と，多義性解消への適用では，クラスタリング時に，教師情報を部分的に利用する，半教師ありクラスタリング手法を開発している．半教師ありクラスタリングでは，用例対が同じ語義に対するものである，あるいは，異なる語義に対するものである，ある用例がある語義に対するものである，等の事実を既知のものとしてシステムに与えることで，より精度の高いクラスタリングを実現する．語義タグ付けの支援において利用できるだけでなく，半教師ありクラスタリングは，以下の点においても利用できると我々は考えている．

_ 新語義候補の検出，

_ 多義性解消システムの性能改善．

従来のクラスタリング手法に比べ，高精度のクラスタリングが実現できることから，より精度の高い新語義候補検出が期待できる．また，用例のクラスタリング結果の情報を利用することで，より性能の良い多義性解消手法が実現できる．

代表性のあるコーパスを用いた語義曖昧性解消では，複数のジャンルのテキストに対する語義タグ付コーパスが徐々に構築できてきており，昨年度語義曖昧性解消における領域適応に関する研究に着手し，ソース(適応元)データとターゲット(適応先)データの性質により，ソースデータとターゲットデータの組み合わせごとに効果的な領域適応手法が異なることが分かっている．そのため，ソースデータとターゲットデータの組み合わせごとに効果的な領域適応手法を自動的に選択する手法の開発を行っている．領域適応手法の自動選択は，ソースデータとターゲットデータの性質に関する情報を元に決定木学習を用いて行う．自動的に選択された領域適応手法を用いることで，語義曖昧性解消の性能が有意に向上することが確認されている．

2) コーパスからの新語義の発見

北陸先端科学技術大学院大学の研究グループでは，コーパスから単語の新しい意味・用法を発見する研究に取り組んできた．

まず，対象単語を含む用例をコーパスから収集し，同じ語義を持つ用例をまとめたクラスタを作成する．用例は以下の4 種類の特徴ベクトルで表現する．対象語の直前または直後に現われる単語を素性とする隣接ベクトル，対象語の周辺に出現する単語を素性とする文脈ベクトル，対象語と二次共起(間接共起) する単語を素性とする連想ベクトル，テキストのトピックを素性とするトピックベクトルである．クラスタリングの際には，これらの特徴ベクトルのいずれかの類似度が高い用例をまとめてクラスタを作成する．1 種類の特徴ベクトルを用いる先行研究と比べ，複数の特徴ベクトルを同時に用いることで，語の類似性を様々な観点から評価できる点に提案手法

の特徴がある．

次に，用例クラスタが新語義を持つ用例の集合であるかを判定する．クラスタ集合をC，辞書で定義されている語義の集合をS とし（NSは新語義を表わす），用例クラスタと語義を対応付けるマッピング関数M： C ！ S を決める．この際，(1) M で対応付けられたクラスタと辞書の語義の類似度が高く，(2) 辞書のどの語義とも類似度が低いクラスタは新語義NS に対応させ，(3) 似ているクラスタは同じ語義に対応付ける場合に高くなるようなマッピング関数のスコアを定義し，それが最大となるM を1つ選択する．選択されたMにおいてNS に対応付けられたクラスタを新語義の用例を集めたクラスタとして出力する．

3) 特異用例の検出

茨城大学の研究グループでは，特異用例の検出に取り組んできた．特異用例とは対象単語の少し変わった用法をもつ用例のことである．このような用例を検出することで，語義識別に対する質の高い訓練データを作成することができる．また特異用例はその対象単語の言語的性質を調べる際にも役立つ．さらに特異用例の存在の有無により，利用しているコーパスの均一性や代表性も評価できる．

提案手法は2つの検出手法からなる．第1 の手法はLOF 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，教師データから語義識別の分類器を学習し，各データの語義を推定する．推定された語義のクラスタとデータとの距離関係から外れ値かどうかを判定する．提案手法では第1 の手法により外れ値の候補を取り出し，第2 の手法でその候補を選別する．

4) 同時クラスタリングを利用した動詞類義語獲得

岡山大学の研究グループでは項構造レベルの動詞辞書を人手で構築するために大規模テキストから自動的に語義を抽出する手法の開発を行ってきた．動詞項構造辞書とは動詞の類語を概念としてシソーラス形式でまとめ，さらに動詞の使用例について項の意味役割まで付与した事例とリンクさせたデータである．既に人手による構築で，4425 語(7473 語義)の動詞に対して例文付きで辞書を構築しているがさらなる拡張を行うために半自動でテキストから辞書知識を構築する手法について検討を行ってきた．その結果，動詞類語を獲得する手法を新たに提案することができた．また現状では意味役割まで付与した事例をテキストから獲得するためのツールおよび評価コーパス開発までおこなっている．動詞類義語獲得では，動詞の類語をテキスト中の係り受け関係から獲得する手法として，同時クラスタリングがベクトルベースの動詞だけのクラスタリングに対して有効であることを明らかにした．

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　６件）

① ▪Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, Hikaru Yokono, On SemEval-2010 Japanese WSD Task, 自然言語処理, Vol.18, No.3, 2011.(査読有)

②Koichi Takeuchi and Hideyuki Takahashi, Co-clustering with Recursive Elimination for Verb Synonym Extraction from Large Text Corpus, IEICE Transactions on Information and Systems, Vol.E92-D, No.12, pp. 2334-2340, 2009.(査読有)

③ Akemi Tera, Kiyoaki Shirai, Takaya Yuizono, Kozo Sugiyama. Analysis of Eye Movements and Linguistic Boundaries in a Text for the Investigation of Japanese

Reading Processes. IEICE Transaction on Information and Systems, Special Issue on Knowledge, Information and Creativity Support System, Vol. E91-D, No. 11, pp.2560-2567, 2008.(査読有)

〔学会発表〕(計 47 件)

① Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, Hikaru Yokono. SemEval-2010 Task: Japanese WSD. The 5th International Workshop on Semantic Evaluation, pp.69--74, 2010 年 7 月 15 日、Uppsala.

② Kiyoaki Shirai, Makoto Nakamura. JAIST: Clustering and Classification Based Approaches for Japanese WSD. The 5th International Workshop on Semantic Evaluation, pp.379-382, 2010 年 7 月 15 日、Uppsala.

③ Hiroyuki Shinnou, Minoru Sasaki、Detection of Peculiar Examples using LOF and One Class SVM, LREC-2010, 2010 年 5 月 21 日, Malta.

④ Minoru Sasaki, Hiroyuki Shinnou, Document Clustering Using Semantic Relationship Between Target Documents And Related Documents, The Fourth International Conference on Advances in Semantic Processing, 2010 年 10 月 27 日, Florence.

⑤ Koichi Takeuchi, Kentaro Inui, Nao Takeuchi and Atsushi Fujita, A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings, The 8th Workshop on Asian Language Resources, 2010 年 8 月 21 日, Beijing.

〔その他〕
ホームページ等

1) BCCWJ を用いた新しい語義曖昧性解消タスク
http://oku-gw.pi.titech.ac.jp/wsd.html

2) 意味役割付与システムの公開
http://cl.cs.okayama-u.ac.jp/study/project/sea.html

3) 動詞の概念辞書の公開
http://cl.cs.okayama-u.ac.jp/rsc/data/index.html

６．研究組織
(1)研究代表者
奥村 学 (OKUMURA MANABU)
東京工業大学・精密工学研究所・教授
研究者番号：60214079

(2)研究分担者
白井 清昭 (SHIRAI KIYOAKI)
北陸先端科学技術大学院大学・情報科学研究科・准教授
研究者番号：30302970

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

高村 大也 (TAKAMURA HIROYA)
東京工業大学・精密工学研究所・准教授
研究者番号：80361773

竹内 孔一 (TAKEUCHI KOUICHI)
岡山大学・自然科学研究科・講師
研究者番号：80311174

佐々木 稔 (SASAKI MINORU)
茨城大学・工学部・講師
研究者番号：60344834

中村 誠 (NAKAMURA MAKOTO)
北陸先端科学技術大学院大学・情報科学研究科・助教
研究者番号：50377438