様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　１　日現在

機関番号：　１２１０１
研究種目：　若手研究(B)
研究期間：　2012～2015
課題番号：　２４７００１３８

研究課題名（和文）訓練事例の最適化による語義曖昧性における領域適応

研究課題名（英文）Domain Adaptation for Word Sense Disambiguation using Optimization of Training Data

研究代表者
古宮　嘉那子（Kanako, Komiya）

茨城大学・工学部・講師

研究者番号：１０５９２３３９

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：語義曖昧性解消の領域適応のための訓練事例の最適化のために，訓練事例集合をいくつも作成した上で，どの訓練事例集合がよいのかを選ぶ手法を採用した．この基準として，教師ありの語義曖昧性解消の領域適応に利用していた，確信度（SVMの長平面からの距離など，active learningで用例選択に利用される尺度）とを利用するのが良いことがわかった．また，LOO-Boundという，SVM に対し Leave-One-Out Estimation を行ったときのエラーの期待値の上限を利用したスコアを併用するとよいことが分かった．

研究成果の概要（英文）：The research project developed the method to optimize the training data set of domain adaptation for word sense disambiguation. Firstly a number of training data set were generated and then the best one is selected using some criteria. The degree of confidence, which is a criterion used for the instance selection when active learning is carried out, and a score using LOO-Bound, which is an error rate of SVM when leave-one-out estimation is performed, are used for the criterion.

研究分野：　自然言語処理

キーワード：　領域適応　語義曖昧性解消　用例選択

１．研究開始当初の背景

語義曖昧性解消で答えとなる語義は、使用するコーパスの分野に大きく依存する。そのため、機械学習を用いて語義曖昧性解消を行う場合には、多量な用例を備えた対象分野（以下、ターゲットドメイン）のラベル付きコーパスを訓練事例に使用するのが最も望ましい。しかし、ターゲットドメインのコーパスが手に入らない場合があり、この場合には、別の分野（ソースドメイン）のコーパスをうまく適応して利用する「領域適応」が必要となる。研究開始当時、自然言語処理において、領域適応の研究が盛んになってきていた。また、おもに構文解析における領域適応の研究で、多数のコーパスの中から最もターゲットドメインのコーパスに近いコーパスを選んで、訓練用コーパスとして利用することにより、タスクの正解率をあげようというコーパス選択の研究が行われていた（Van Ash, ACL2010）。

また、単一のソースドメインのコーパスを細切れにしてサブコーパスをいくつも作り、それらのサブコーパスを小さなコーパスとみなして、そのうち最もターゲットドメインのコーパスに近いと思われるものを訓練用コーパスとして利用した研究もあった。（Axelrod et al., EMNLP2011）.

その一方で、ターゲットドメインのデータとソースドメインのデータの組み合わせによって、最適な領域適応手法は異なる（Komiya，Okumura，IJCNLP2011）。また、この際，ターゲットドメインのデータとソースドメインのデータ間の素性分布の距離が、使うべき手法を分ける大きな手掛かりとなっている。

２．研究の目的

本研究では、多数の入手可能なコーパスから得られる用例を適切に選択し、対象分野のコーパスから得られる事例ベクトル集合に最も近くなるように、訓練事例ベクトル集合を最適化する手法を研究する。その際、ターゲットドメインのラベルなしコーパスから得られる事例ベクトル集合の素性分布にできるだけ近い素性分布をもつように、訓練用の事例ベクトル集合を最適化する。こうすることで、ラベル付きコーパスが手に入らない際の語義曖昧性解消において、ターゲットデータとソースデータに対し、高い正解率を与えうる、最適な訓練事例ベクトル集合を用意することができるようになる。そのためには、語義曖昧性解消において事例ベクトル集合の間の類似度を測ることができる類似度を定義することが必要である。パープレキシティ、JS 距離をはじめ、さまざまな観点からの距離尺度によって最もターゲットドメインのコーパスから得られる事例ベクトル集合に「似ている」訓練用の事例ベクトル集合を作成し、それぞれの類似度の性能について比較する。つまり、申請者は、語義曖昧性解消の領域適応において、

1)多数の入手可能なコーパスから、訓練事例ベクトルを作成し、
2) 事例ベクトル集合間の類似度を定義して、
3)対象分野のコーパスから得られる事例ベクトル集合に最も近い訓練事例ベクトル集合になるよう、訓練事例ベクトル集合を最適化する手法について研究を行う。

３．研究の方法

平成 24 年度は、本研究の関連研究を調査するとともに、コーパス収集、事例ベクトル集合の作成、基本的な類似度を基準に用いた訓練事例ベクトル集合の最適化といった、基本的な一連の流れを実装し、訓練事例ベクトル集合の最適化によって、語義曖昧性解消の正解率が上がるかどうかを調べた。

また，最適化において基準とする類似度のうち、より語義曖昧性解消の正解率を高める類似度は何なのかを調べるため、さまざまな類似度を基準とした訓練事例ベクトル集合の最適化を実装して、結果を比較した．また，教師なしの領域適応においても分類器の確信度によって用例が選べることが分かったため，分類器の確信度を利用した教師なしの領域適応に関する実験を行った．この際，L00-Bound という値も利用した．

平成 25 年度は、調査した関連研究と平成 24 年度の結果を踏まえ、確信度の調整を行った．特に，訓練事例数が少ない際に，確信度の信頼性が低いことに着目し，そのための改良を行った．

平成 26 年度は、平成 25 年度までの結果について考察し、訓練事例を反復的に選択することを許すことで，よりよい領域適応の訓練事例を作成することを目指した．この際，訓練事例を増やして新しいデータセットで実験を行った．また，これまでの知見を生かし，語義曖昧性解消の領域適応だけではなく，固有表現抽出に関する領域適応の実験を行った．

平成 27 年度は，分散表現を利用した語義曖昧性解消の領域適応の実験を行うとともに，これまでの成果についての発表を行った．

４．研究成果

語義曖昧性解消の領域適応のために，訓練事例を最適化する研究を行った．本節では，副次的に得られた成果（固有表現抽出における領域適応や，語義曖昧性解消についての別手法を利用した領域適応，訓練事例選択の際

に思いついた文書分類の新手法など）については触れず，もともと想定していたタスクのみについて触れる．なお，後述される成果は，これらの副次的な研究成果も全て含めたものである．

語義曖昧性解消の領域適応のための訓練事例の最適化の研究は，具体的には，訓練事例集合をいくつも作成した上で，どの訓練事例集合がよいのかを選ぶ手法を採用した．

まず，さまざまな類似度（ユークリッド距離（ED），コサイン類似度（CS），ジャッカード係数（JSD），ダイス係数（DSC），シンプソン係数（SSC），ランド類似度（RS）など）を利用して訓練事例集合を選択することを試してみたが，結果として，それらは教師無しの語義曖昧性解消の領域適応には適していないことが分かった．一方で，教師ありの語義曖昧性解消の領域適応に利用していた，確信度（SVM の長平面からの距離など，active learning で用例選択に利用される尺度）を利用するのが良いことがわかった．

次に，この確信度を利用して，ふたつの実験を行った．ひとつは，この確信度を利用して，みっつ以上のコーパスから合議を行うことによってよりよい訓練事例集合を求める実験である．この際，

(1)最も高い確信度の分類器の結果（語義）を採用する
(2) 語義ごとに，複数分類器から出力された確信度を積算し，最も高い確信度となった語義を採用する
(3)語義ごとに，複数分類器から出力された確信度を足しあわせ，最も高い確信度となった語義を採用する
(4)分類器ごとに，最も高い確信度となった語義に一票入れ，最も多数の票が入った語義を採用する

の四つを試し，(1)の手法が最もよいことが分かった．また，この際には，訓練事例集合は，用例ごとに定めた．

もうひとつは，複数のコーパスから確信度だけではなく，LOO-Bound という，SVM に対し Leave-One-Out Estimation を行ったときのエラーの期待値の上限を利用したスコアを併用することで，よりよい結果を得られるようにする実験である．このときにはコーパス＝訓練事例集合とはせず，複数のコーパスから用例を混合して持ってきて，訓練事例集合の候補を作成した．この際，確信度は，LOO-Bound と併用することでより良い訓練集合を選出できることが分かった．また，この実験では，訓練事例集合は単語タイプごとに選択した．

しかし，本研究では，訓練事例の事例数についても未知であり，その結果，自動的に選んだ訓練事例集合の訓練事例数が小さいときには，確信度の信頼性が低いことが問題となった．たとえば，極端な例を挙げると，訓練事例集合にひとつの用例しか含まれなかった場合，可能なラベル（システムが答える正解）は自動的にひとつになり，分類問題として考えた場合，確信度は 100%となってしまう．しかし，ひとつの用例しか含まない訓練事例集合が領域適応に最もふさわしいということはたいていの場合あり得ないので，調整が必要となった．そのため，

(1)分類器が出力した語義の，訓練事例における事前 確率でスコアを割る，
(2) 分類器の訓練事例における最頻出語義の事前確率 でスコアを割る，
(3) 分類器の訓練事例に出現する語義数をスコアにかける，

の三つの手法を試した．その結果，(3)の結果が最もよく，調整前に比べて有意な正解率の上昇が見られた．

次に，訓練事例集合の大きさが大きいときには，大抵語義曖昧性解消の結果が良くなることに注目し，反復的に訓練事例集合を追加することで最適な訓練事例集合を選ぶ手法について実験した．残念ながら，この手法は最もよい成果をあげる，ということはできなかったが，反復的な手法により，短い時間である程度の精度の結果をあげることができた．また，この実験でも，訓練事例集合は単語タイプごとに選択した．また、この実験では、使用するコーパスの用例数を増やして実験を行った。

さらに、これまでの研究で、訓練事例集合を、用例ごとに選択する場合と、単語タイプごとに選択する場合の二つの設定があり、どちらの方が優れているのかを実験した。その結果、一用例を一つの重みとして計算する平均であるマイクロ平均は、用例ごとに訓練事例集合を選択したほうが高いが、一単語タイプを一つの重みとして計算する平均であるマイクロ平均は、単語タイプごとに訓練事例を選択したほうが高いことが分かった。

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 新納浩幸，村田真樹，白井清昭，福本文代，藤田早苗，佐々木稔，古宮嘉那子，乾孝司，クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け，自然言語処理．Vo.22．No.5，pp.319-362，(2015.12)．査読有
② 古宮嘉那子，奥村学，語義曖昧性解消の

ための領域適応手法の決定木学習による自動選択，自然言語処理，Vol.19，No.3，pp.143-166，(2012.9)．査読有

〔学会発表〕（計　60　件）

① Kanako KOMIYA，Daichi EDAMURA，Ryuta TAMURA，Minoru SASAKI，Hiroyuki SHINNOU and Yoshiyuki KOTANI，Domain Adaptation with Filtering for Named Entity Extraction of Japanese Anime-Related Words，RANLP 2015，pp.291-297，Hissar，Bulgaria，(2015.09.07)．

② 小林 優稀，古宮 嘉那子，佐々木 稔，新納 浩幸，奥村 学,領域適応のためのサポートベクトルを用いた訓練事例の反復的選択，第七回コーパス日本語学ワークショップ予稿集，pp．129-136，立川，(2015,03,10)．

③ 古宮嘉那子，小谷善行，奥村学，合議による語義曖昧性解消の領域適応のための確信度の調整，第二十回言語処理学会年次大会予稿集，　pp．520-523，札幌，(2014,03,19)．

④ 古宮嘉那子，小谷善行，奥村学，語義曖昧性解消の領域適応のための訓練事例集合の選択,第十九回言語処理学会年次大会予稿集，pp.940-943，名古屋，(2013,03,15)．

⑤ 古宮嘉那子、奥村 学、小谷 善行，分類器の確信度を用いた合議制による語義曖昧性解消の unsupervised な領域適応，第三回コーパス日本語学ワークショップ予稿集，pp．1-6，立川，(2013,02,28)．

⑥ 堀内 浩史郎、古宮嘉那子、小谷 善行，語義曖昧性解消の領域適応のための訓練データの選択法 ～複数ドメインからの選択～，第三回コーパス日本語学ワークショップ予稿集，pp．97-102，立川，(2013,02,28)．

⑦ Kanako KOMIYA，Manabu OKUMURA．Automatic Domain Adaptation for Word Sense Disambiguation Based on Comparison of Multiple Classifiers，PACLIC 2012，pp 77-85，Bali，Indonesia，(2012.11.08)．

〔図書〕（計　1　件）

６．研究組織

(1)研究代表者

古宮　嘉那子（Kanako Komiya）

茨城大学　情報工学科　講師

研究者番号：10592339

(2)研究分担者

なし

(3)連携研究者

なし