半教師あり語義曖昧性解消における各ジャンルの語義なし用例文の利用

半教師あり語義曖昧性解消における各ジャンルの語義なし用例文の利用

谷田部 梨恵（茨城大学大学院理工学研究科）†
佐々木 稔（茨城大学工学部情報工学科）

Semi-Supervised Word Sense Disambiguation Using Unlabeled Examples of Each Genre

Rie Yatabe (Ibaraki University)
Minoru Sasaki (Ibaraki University)

要旨

単語の語義曖昧性解消は，今日に至るまで様々な研究が行われており，教師あり学習を用いることで高い精度を出している．しかし，先行研究では学習用のデータが不足して誤る事例が多いことが指摘されている．新たに学習データを追加するには，用例文における単語の正解語義の割り当てに精通した専門家によるラベル付与が必要となるためコストがかかるという問題がある．この問題を解決するために，グラフベースの半教師あり学習を用いた語義曖昧性解消を提案し，語義なし用例文の利用による精度改善を行う．そこで，BCCWJ の各ジャンルにおける語義なし用例文に対して語義曖昧性解消精度の比較を行い，どのような語義なしデータの利用が有効なのか分析を行う．実験の結果，BCCWJ 全ての用例文を追加した場合よりも精度が低くなったが，今回扱ったジャンルの中では雑誌（PM）に含まれる用例文を追加した場合が最も高い精度結果となった．そのため，ジャンルを限定して語義なし用例文を追加しても，語義曖昧性解消の精度にあまり効果がないことが明らかとなった．しかし，教師あり学習との語義曖昧性解消精度との比較を行った結果，グラフベースの半教師あり学習の語義曖昧性解消精度が高くなったため，グラフベースの半教師あり学習は学習データ不足の改善に有効であると考えられる．

１．はじめに

文章中で使われる単語の語義曖昧性解消は，今日に至るまで様々な研究が行われており，教師あり学習である Support Vector Machine(SVM)では高い精度を出している．更に精度を高めることを目的として，先行研究では，この SVM を使用したシステムの誤り原因の分類が行われ，学習用のデータが不足して誤る事例の多いことが指摘されている（新納，村田，白井，福本，藤田，佐々木，古宮，乾 2015）．しかし，新たに学習データを追加するには，用例文における単語の正解語義の割り当てに精通した専門家によるラベル付与が必要となるためコストがかかるという問題がある．

そこで，本研究ではグラフベースの半教師あり学習を用いた語義曖昧性解消を提案し，語義なし用例文の利用による精度改善を行う．本稿では，書籍・雑誌・白書・Yahoo!知恵袋の4つのジャンルの日本語のコーパスに対し，ジャンルごとに語義曖昧性解消精度の比較を行う．BCCWJ の各ジャンル結果の比較を行い，どの文書のジャンルの利用が有効なのか分析を行うことを目的とする．

２．関連研究

これまでの研究において，半教師あり学習手法を利用した様々な分類の方法が提案されている．代表的な手法として，ラベル付きデータから作成した分類器の予測結果に基づく手法やデータをある空間へマッピングする手法が存在する．前述の手法は Co-training や Self-training がある．

† 19nm732r@vc.ibaraki.ac.jp

Co-training や Self-training による分類手法はラベル付きデータから得られる分類器を使用し，ラベルなしデータに確信度付きのラベルを付与して，それを利用することで分類器を改善し，その上で学習と識別行う．後述のデータをある空間へマッピングする分類手法は多様体論を応用した手法や生成モデル，Variational Autoencoder(VAE) を使用した手法が含まれる．これらの手法はまずラベルなしデータを分離し，空間にマップする． 次にラベル付きデータもその空間にマップし，その空間上で分類器の学習と識別を行う．この他に，ラベル付きデータの素性に一致するデータは同じラベルであると仮定してラベルを付与する藤田ら(2011)の手法も存在する．

グラフ構造に基づいてラベルを予測する手法は前述のラベル付きデータから作成した分類器の予測結果に基づく手法のうちに入る．これと関連した手法として，ラベル伝搬法(LP)がある．これは用例文から抽出した素性データのグラフを作成し，ラベルの自動推定を行う手法である．類似度の最も高い文同士は同一語義を持つと仮定し，ラベルを伝搬させることでラベル無しデータにラベルを付与する．本研究では，グラフを基づいてラベルを予測する手法をとるが，ラベル付きデータとグラフ構造の学習を同時に行う手法となっている．

３．半教師あり学習手法

本節では，グラフベースの半教師あり学習手法と語義曖昧性解消のシステムについて述べる．

３.１　グラフベースの半教師あり学習：Planetoid

Planetoid は半教師あり深層学習手法として 2016 年に提案されたものである．この手法では，訓練データとラベルなしデータの集合と用例文間の関係を表すグラフを入力し入力データから学習と推論を行う(Zhilin，William and Ruslan 2016)．

Planetoid の簡易的なネットワーク構造(Zhilin，William and Ruslan 2016)を図 1 に示す．図 1 のネットワーク構造を利用し，グラフ構造と訓練データを同時に学習させる．損失関数では訓練データを学習したときの損失とラベル無しデータでグラフ構造を予測したときの損失の二つの合計を最小化させる．損失に応じてフィードバック学習を行うので，一定回数繰り返し学習させる．

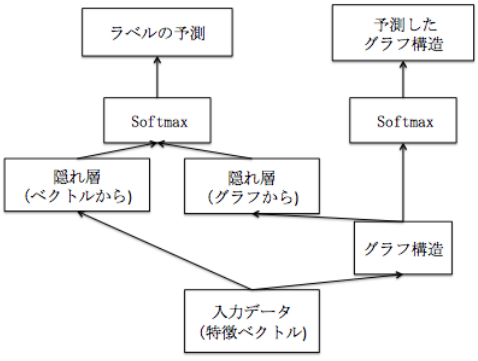

図 1:Planetoid のネットワーク構造

３.２　語義曖昧性解消システムの概要

本システムは語義識別モデルの学習と語義を知りたい用例文の語義推定を行う．語義識別モデルは訓練データと文の関係を示すグラフを入力して学習を行うことで得られる．流れは学習用データのベクトル化を行い，得られたベクトルからグラフ構造を作成する．訓練データのベクトルと先ほど得たグラフ構造を同時に学習させ，識別モデルを得る．そこで得られた語義識別モデルに用例文を入力することで対象単語の語義を推定することが可能となる．

３.３　データの前処理

データ入力部分では教師データと語義なし用例文，テストデータ共に下記に示す方法によりベクトル化を行う．まず対象単語を含む用例文に対して形態素解析を行い，対象単語及び前後二単語の単語，品詞，品詞大分類，係り受け，シソーラス情報を素性として抽出する．その後，これらの素性に対する出現頻度を割り当てることで，用例文をベクトル化する．この作業を教師データ，テストデータは 50 文ずつ，語義なし用例文は「日本語書き言葉均衡コーパス(BCCWJ)」である白書・書籍・雑誌・Yahoo!知恵袋からそれぞれ対象単語を含む用例文を抽出し，ベクトル化する．本稿では，形態素解析ツールとして日本語形態素解析システムは「UniDic[1]」を使用する．

３.４　入力するグラフ構造

訓練データとラベル無しデータに対して，Planetoid に入力するためのグラフ構造を作成する．グラフ構造は訓練データとラベル無しデータに含まれる各用例文をノードとし，ノード間の類似度をエッジとする．ノード間の類似度は訓練データとラベル無しデータのベクトルを用いて計算する．各ノードからエッジを張るノードは，最も類似度の高いノードとしきい値以上の類似度を持つノードとする．エッジをつなげる様子を図 2 に示す．

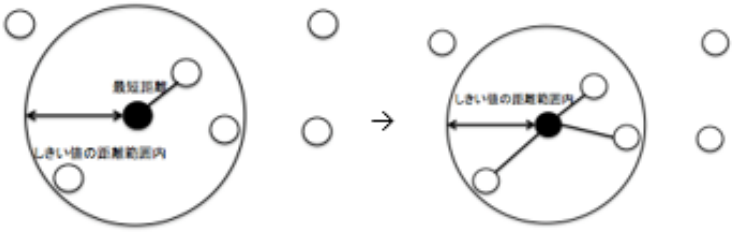

図 2:グラフにおけるエッジのつなぎ方

本稿ではノード間の類似度計算手法として Jaccard 係数を使用する．Jaccard 係数は二つの集合間で共通する単語の数の比率を求める．一文に含まれる単語ベクトルの集合 A と B が与えられた場合，一致する要素の比率 Jを表す．

$$J(A,B)=|A\cap B|/|A\cup B|, \quad (0\leq J(A,B)\leq 1)$$

３.５　Planetoid を用いた学習手法

学習手法にはミニバッチの確率的勾配降下法(SGD)を使用している(Bottou, 2010). この学習法は学習用データ(訓練データとグラフ)の中から，いくつかデータを取り出して損失関数を計算し，最適化することでモデルパラメータwを更新する．損失関数 L(w)と学習率 ε を用いて以下の式である勾配ステップをとることで最適なモデルパラメータをとる．損失関数を最適化したことで得られた語義識別器を次節の識別方法で使用する．

$$w = w - \epsilon \left(\partial L(w) / \partial w \right)$$

３.６　識別方法

前節で得られた識別器にベクトル化したテストデータを入力することで，自動識別した語義を出力する．出力した語義と正解の語義を比較して正解率を求める．

４．実験

本節では，グラフベースの半教師あり学習による BCCWJ の各ジャンルにおけるラベルな

[1] https://unidic.ninjal.ac.jp/download

し用例文に対して語義曖昧性解消を行い，どのようなラベルなしデータの利用が有効なのか調査するために，BCCWJ の各ジャンルの精度比較実験を行う．また，語義曖昧性解消精度の最も高いジャンルと SVM の精度比較実験を行う．

４.１　実験データ

本研究における対象単語は，Semeval2010 日本語 WSD タスクデータである対象単語の 50 個を利用する(Okumura, Shirai, Komiya, Yokono, 2010)．また，訓練データとテストデータはその単語を使用した用例文の文章データをそれぞれ 50 個用意されている．それぞれの領域で学習を行い，対象単語の意味を調査する．

実験用の語義なし用例文データには，国立国語研究所が開発した現代日本語書き言葉均衡コーパス(BCCWJ)を利用する．BCCWJ は日本語の様々なジャンルの文書を収録した，書き言葉の全体像を把握するために構築されたコーパスである．今回の実験では「書籍や雑誌」(以下，PB)「Yahoo! 知恵袋」(OC)「白書」(OW)「雑誌」(PM)の 4 つのジャンルに含まれる文書データを使用する．また，ジャンルごとに用例文数が異なり，用例文の多さとしては PB＞OC＞OW＞PM の順になっている．

４.２　実験の設定

この実験ではグラフを作成するとき，学習用データのノードに対し最短距離ノードを加え，さらに Jaccard 係数が 0.9 以上の類似度(同じ語義であるという確信度が高い)を持つデータをノードとする．ここで，同じ語義を持つ用例文は周辺に類似した単語や品詞が出やすく，異なる語義を持つ用例文は周辺に異なった単語や品詞などが出現しやすいと仮定している為，最高類似度だけでなく，確信度の高い用例文も組み込む事が重要であると考えたため，このような設定とした．

今回の事前学習において訓練データの学習は 10000 回，グラフ構造の学習は 1000 回行っている．事前学習で得られた初期値を用いた学習とテストデータの識別を 1000 回繰り返すことで語義を推測する．

５．実験結果

グラフベースの半教師あり学習による，BCCWJ の各ジャンルと全てのジャンルにおける語義なし用例文を追加した場合に対する実験結果を表 1 に示す．表 1 の結果を見ると，各ジャンルの精度では PM の語義曖昧性解消精度が最も高くなったが，BCCWJ 全ての文書追加の結果より低くなった．また，PB や OC は追加可能な用例文の数が OW や PM よりも多かったにも関わらず，語義曖昧性解消の精度が低くなった．

表 1:各ジャンルと BCCWJ 全て追加の語義曖昧性解消精度

データのジャンル	50 単語の平均精度
PB	75.88%
OC	75.76%
OW	76.52%
PM	76.88%
BCCWJ	77.76%

表 2：PM と SVM の語義曖昧性解消精度

対象単語	PM	SVM	対象単語	PM	SVM
117　相手	80%	84%	34522　強い	94%	92%
166　会う	90%	88%	34626　手	78%	78%
545　上げる	60%	60%	35478　出る	60%	56%
755　与える	70%	70%	35881　電話	80%	80%
1889　生きる	94%	94%	37713　取る	28%	28%
2843　意味	46%	48%	40289　乗る	70%	72%
2998　入れる	72%	76%	40333　場合	86%	88%
5167　大きい	98%	94%	40699　入る	66%	66%
5541　教える	40%	52%	41135　はじめ	96%	96%
8783　可能	62%	48%	41138　始める	90%	88%
9590　考える	98%	98%	41150　場所	96%	94%
9667　関係	96%	96%	41912　早い	70%	70%
10703　技術	84%	82%	43494　一	94%	92%
14411　経済	98%	98%	44126　開く	84%	84%
15615　現場	74%	74%	46086　文化	98%	98%
17877　子供	68%	64%	47634　他	100%	100%
20676　時間	82%	82%	48488　前	84%	76%
21128　市場	58%	60%	49355　見える	70%	70%
22293　社会	86%	84%	49812　認める	76%	76%
24646　情報	82%	82%	50038　見る	82%	72%
26839　進める	62%	92%	51332　持つ	86%	86%
27236　する	66%	78%	51409　求める	70%	64%
31166　高い	86%	86%	51421　もの	88%	88%
31472　出す	42%	36%	52310　やる	96%	96%
31640　立つ	56%	58%	52935　良い	52%	46%
			50 単語の平均精度	76.88%	76.8%

次にジャンル別で最も語義曖昧性解消精度の高かった PM の詳細結果と訓練データとテストデータのみを使用した SVM の詳細結果を表 2 に示す．表 2 の結果を見ると，PM の 50 単語の平均精度が SVM より高くなっている．SVM より精度が高くなった単語は 16 個あり，中でも「可能」と「見る」は 10%以上精度が向上している．SVM より精度が下がってしまった単語は 10 個あり，中でも「進める」,「する」や「教える」は 10%以上精度が下がった．

６．考察

実験結果より，各ジャンルの精度比較では PM が最も高い精度を出したが，BCCWJ 全ての文書追加した結果より低くなったため，ジャンルを限定して語義なし用例文を追加しても語義曖昧性解消の精度に効果が得られないと考えられる．PM は PB や OC，OW のジ

ャンルより追加できる用例文が少ないにも関わらず，精度が高かった．そのため，追加する用例文数の観点からみると，用例文数が異なっても語義曖昧性解消精度にそれほど変化がないと考えられる．また，表 2 の実験結果から PM の 50 単語の平均精度が SVM より高くなっているため, PM を使用した半教師あり語義曖昧性解消は学習データ不足の改善に有効であると考えられる．

PM の次に精度の高かった OW は偏った内容であることから，訓練データやテストデータと近い文が少なくなったと考えられる．そのため，OW の用例文を追加した場合は語義曖昧性解消の精度が低くなる可能性がある．PB や OC は追加できる用例文が多いにも関わらず精度が低かったので，様々な種類の用例文を追加することより，訓練データやテストデータに近い用例文を集めて追加すると効果が高いのではないかと考える．

7．結論

本稿では，グラフベースの半教師あり学習を用いた語義曖昧性解消を利用し，BCCWJ の各ジャンルにおける語義なし用例文に対して語義曖昧性解消精度の比較を行い，どのような語義なし用例文の利用が有効なのか分析を行った．実験の結果，BCCWJ 全ての用例文を追加した場合よりも精度が低くなったが，今回扱ったジャンルの中では雑誌（PM）に含まれる用例文を追加した場合が最も高い精度結果となった．そのため，ジャンルを限定して語義なし用例文を追加しても，語義曖昧性解消の精度にあまり効果がないことが明らかとなった．また，PM は他のジャンルより追加できる用例文が少ないにも関わらず，精度が高かったため，用例文数が異なっても語義曖昧性解消精度にそれほど変化がないことが示された．そして，白書（OW）のように内容が偏ると精度が低くなる可能性があるのではないかと考えた．これより，各ジャンルの語義なし用例文を追加した実験結果から，追加する用例文の数に関係せずに訓練データやテストデータに近いデータを追加した方が精度を高めるのではないかと考えられる. SVM との語義曖昧性解消精度の比較を行った結果では PM を使用したグラフベースの半教師あり語義曖昧性解消の精度が高くなったため，グラフベースの半教師あり学習は学習データ不足の改善に有効であると考えられる．

今後は，今回扱った以外のジャンルでも実験を行いさらなる分析を行うことやデータから訓練データやテストデータに近い語義なし用例文を追加することで精度が向上するかどうか確認することが課題である．

文　献

Bottou, L.　(2010). “Large-scale machine learning with stochastic gradient descent.” In *COMPSTAT*, pp. 177–186.

藤田早苗，Kevin Duh，藤野昭典，平博順，進藤裕之 (2011). 「日本語語義曖昧性解消のための訓練データの自動拡張」自然言語処理, 18(3),pp.273-291

Okumura, M., Shirai, K., Komiya, K., Yokono, H. (2010). “Semeval-2010 task: Japanese WSD.” In: Proceedings of the SemEval-2010, ACL 2010, pp. 69–74

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　(2015). 「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, 22 (5), pp.319-362

Zhilin Yang，William W. Cohen ，Ruslan Salakhutdinov (2016). “Revisiting Semi-Supervised Learning with Graph Embeddings” In ICML 2016，volume 48，pp40-48

関連 URL

『現代日本語書き言葉均衡コーパス』　http://pj.ninjal.ac.jp/corpus_center/bccwj/