Vol.2019-NL-241 No.8
2019/8/29

グラフニューラルネットワークを用いた半教師あり語義曖昧性解消

谷田部梨恵[†1(a)]　佐々木稔[†2(b)]

概要：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．先行研究では，このアプローチにおける識別誤りの主要な要因として学習用のデータ不足を挙げている．そのため，精度を向上するためにはさらに多くの用例文の追加が求められている．しかし，学習用のデータを新たに追加することは，語義識別に精通した専門家による正解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，本研究ではグラフニューラルネットワークを用いた半教師あり語義曖昧性解消手法を提案し，提案手法が語義識別精度の改善に有効であるか分析を行う．実験の結果，教師あり学習の結果を超えることができたので，学習データ不足の改善が示された．そのため，グラフニューラルネットワークを用いた半教師あり語義曖昧性解消手法は語義識別精度の改善に有効であることが示された．また，今回の実験方法では，形態素解析の辞書に「UniDic」を使用することや類似度計算に Jaccard 係数を使用することが語義曖昧性解消精度の向上に効果的だと示された．しかし，既存研究における半教師あり学習の語義曖昧性解消精度の結果を超えることができなかった．ただし，入力データの点で提案手法は既存研究の半教師あり学習と比べると，類似度計算をする事で比較的簡単にラベル無しデータを追加する事が可能なので今後の有効性は高いと考えられる．

Semi-supervised Word Sense Disambiguation Using Graph Convolutional Neural Network

RIE YATABE[†1(a)]　MINORU SASAKI[†2(b)]

1. はじめに

私たちの日常で使用する単語には様々な意味を持つものがある．例えば「意味」という単語では，「その言葉の表す内容．意義.」と「表現や行為の意図・動機」などという語義[1]を複数持つように，単語には様々な語義が存在する．このように様々な語義を持つ単語に対して，前後の文などを参考にして適切な語義を判定する語義曖昧性解消というものがある．

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)[2]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[3]．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，本研究ではグラフニューラルネットワークを用いた半教師あり語義曖昧性解消手法を提案し，提案手法が語義識別精度の改善に有効なのか分析を行うことを目的とする．

†1 茨城大学大学院理工学研究科情報工学専攻
Ibaraki University
†2 茨城大学工学部情報工学科
Ibaraki University
(a) 19nm732r@vc.ibaraki.ac.jp
(b) minoru.sasaki.01@vc.ibaraki.ac.jp

2. 関連研究

語義曖昧性解消を自動的に行う場合，対象となる多義語に対して複数ある語義候補から適切な語義を選択する分類問題として定式化される．そのアプローチは大きく分けて，知識に基づく手法，教師あり学習，教師なし学習，および，半教師あり学習を用いた手法が存在する．本研究は半教師あり学習に基づくアプローチを採用している．

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語義曖昧性解消手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[4]．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[5]や生成モデル[6]，Stacked Denoising Autoencoder を使用した手法[7]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の手法も存在する[8].

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)[9]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，グラフニューラルネットワークを用いた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 グラフニューラルネットワークを用いた半教師あり学習：（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[10]．

Planetoidの簡易的なネットワーク構造を図 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．

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，岩波国語辞典の例文を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,品詞細分類（UniDic は品詞大分類），係り受け，シソーラス情報を素性として抽出する．この詳細なデータ抽出は以下の 20 種類(e1～e20)の素性となっている．

e1=二つ前の単語，e2=二つ前の品詞，e3=その細分類，
e4=一つ前の単語，e5=一つ前の品詞，e6=その細分類，

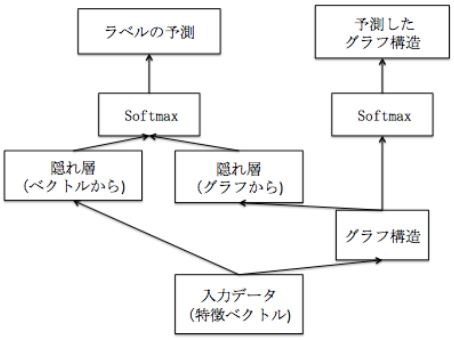

図 1　Planetoid のネットワーク構造

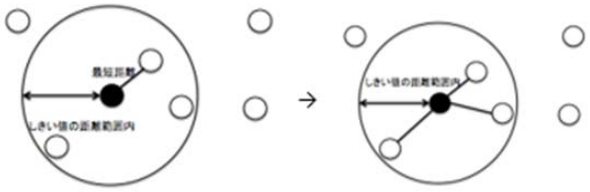

図 2　文グラフにおけるエッジのつなぎ方

e7=対象の単語，e8=対象単語の品詞，e9=その細分類，
e10=一つ後の単語，e11=一つ後の品詞，e12=その細分類，
e13=二つ後の単語，e14=二つ後の品詞，e15=その細分類，
e16=係り受け，e17=二つ前の分類語彙表の値，e18=一つ前の分類語彙表の値，e19=一つ後の分類語彙表の値，e20=二つ後の分類語彙表の値

分類語彙表の ID は 5 桁のものを使用している．また，一つの単語に対して分類語彙表 ID は複数存在するため，e17~e20 に対する素性は複数存在する．

以上の 20 種類の素性に加えて対象単語及び前後三語以内の bigram，trigram，skipbigram を追加した素性(e21～e55)も抽出する．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，岩波国語辞典は各対象単語の例文，語義なし用例文は「日本語書き言葉均衡コーパス(BCCWJ)」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]」と「ipadic[2]」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

1 https://unidic.ninjal.ac.jp/download
2 http://taku910.github.io/mecab/

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．

本稿ではノード間の類似度計算手法として Jaccard 係数 J を使用する．Jaccard 係数 J 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．

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

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[11]．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．

$$w = w - \varepsilon(\partial L(w) / \partial w)$$

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，グラフニューラルネットワークを用いた半教師あり学習による語義曖昧性解消を行い，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[12]．また，訓練データとテストデータとして，その単語を使用した用例文データがそれぞれ 50 個用意されている．それぞれの語義曖昧性解消手法を用いて学習を行い，対象単語の意味を自動的に識別する．

岩波国語辞典の例文の追加は藤田らの抽出方法[8]を参考にして行った．抽出したデータは訓練データに追加して実験を行う．

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．

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 係数は 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．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 回行った．事前学習で得られた初期値を用いた学習を 1000 回繰り返した後で，テストデータの語義識別を行うことで語義を推測する．

SVM の語義曖昧性解消精度比較だけでなく，既存研究の半教師あり学習を用いた語義曖昧性解消精度比較を行うため，藤田らの方法で用いられた最大エントロピーモデル（ME）に訓練データを自動拡張した方法[8]を再現した実験も行っている．

表 1　語義曖昧性解消の精度結果

素性	半教師あり NN	SVM	ME
ipadic + e1～e20	77.24	77.28	-
UniDic + e1～e20	**77.76**	76.8	76.56
UniDic + e1～e20 ＋ 岩波国語辞典	76.68	**77.84**	**76.76**
UniDic + e1～e20 + e21～e55	75.88	75.72	74.92
UniDic + e1～e20 + e21～e55 ＋ 岩波国語辞典	76.28	77.36	76.52

表 2　類似度計算方法を変更したときの実験結果

Jaccard 係数	cosine 類似度
77.76	77.24

表 3　半教師あり NN と ME と文献[8]における識別精度

半教師あり NN	ME	文献[8]
77.76	76.52	79.2

5. 実験結果

グラフニューラルネットワークを用いた半教師あり学習による語義曖昧性解消と SVM と再現実験を行った ME の結果を表 1　語義曖昧性解消の精度結果に示す．また，各手法の最高類似度を太字で表
す．表 1　語義曖昧性解消の精度結果の結果を見ると，半教師あり NN は形態素解析ツールが「UniDic」で e1~e20 までの 20 種類の素性を使用している精度が最も高くなった．また，「UniDic + e1～e20」で半教師あり NN と SVM を比較すると，半教師あり NN の方が高い精度結果となっている．しかし，訓練データに岩波国語辞典の例文を追加した結果では，SVM より精度が低い結果となっている．

半教師あり NN で「UniDic + e1～e20」の素性データを使用し，類似度計算を変更した結果を表 2 に示す．表 2 の結

表 4　半教師あり NN と再現実験の詳細結果

対象単語	ME	半教師あり NN	対象単語	ME	半教師あり NN
117　相手	82	78	34522　強い	92	92
166　会う	90	90	34626　手	68	78
545　上げる	60	58	35478　出る	56	58
755　与える	64	68	35881　電話	78	78
1889　生きる	92	94	37713　取る	44	36
2843　意味	60	44	40289　乗る	50	68
2998　入れる	76	72	40333　場合	72	86
5167　大きい	94	98	40699　入る	70	64
5541　教える	42	56	41135　はじめ	94	96
8783　可能	54	60	41138　始める	84	86
9590　考える	98	98	41150　場所	96	96
9667　関係	90	96	41912　早い	78	70
10703　技術	84	82	43494　一	92	94
14411　経済	98	98	44126　開く	88	84
15615　現場	78	84	46086　文化	90	96
17877　子供	54	64	47634　他	100	100
20676　時間	86	78	48488　前	78	84
21128　市場	70	60	49355　見える	70	68
22293　社会	86	86	49812　認める	80	78
24646　情報	82	82	50038　見る	80	86
26839　進める	70	88	51332　持つ	70	82
27236　する	74	66	51409　求める	76	74
31166　高い	88	86	51421　もの	80	88
31472　出す	66	46	52310　やる	98	96
31640　立つ	30	60	52935　良い	74	58
			50 単語の平均精度	76.52	77.76

果を見ると，Jaccard 係数の精度が cosine 類似度より 0.56%精度が高くなる結果となった．

次に，実験結果で最も高い精度と藤田らの手法を再現した実験結果と文献[8]の実験結果を表 3 に示す．また，半教師あり NN と再現実験を行った ME の詳細な結果を表 4 に示す. 表 3 の結果を見ると, 50 単語全ての平均精度は 1.44%の差があり，既存研究の半教師あり語義曖昧性解消精度の結果は超えられなかった．表 4 の結果では半教師あり NN は ME より精度が高い結果となった．

6. 考察

実験結果より，提案手法は教師あり WSD より精度が向上した．この結果が得られたのは，教師あり WSD と同じ素性リストに加えてラベル無しデータとグラフ構造を用いたことが大きな要因である．そのため，提案手法のシステムによって課題であった学習データ不足に対して改善が見られたと考えられる．

半教師あり NN の訓練データに岩波国語辞典の例文を追加した結果では，追加する前より語義曖昧性解消精度が低くなった．これは岩波国語辞典の例文が短すぎたため，前後の文脈を参考にしてグラフを作成することが出来なかったことや，短い例文の素性が一部一致するだけで，提案手法が使用しているグラフ構造は同じ語義であるとしてしまうため，精度向上の効果が見られなかったのではないかと考えられる．ただし，岩波国語辞典の例文を追加した SVM は効果が見られたので，今後は岩波国語辞典の例文をグラフ構造に悪い影響を与えないように追加する方法を開発することが課題となる．

訓練データに岩波国語辞典の例文を追加し，全ての素性データに前後三語以内の bigram，trigram，skipbigram を追加した結果では，データ追加後の方が下がった．また，この追加方法は SVM も精度が下がっており，素性の追加に効果が見られなかった．これは岩波国語辞典の例文が短いために，前後三語の素性が取れず，正しく語義の分類ができなかったと考えられる．

形態素解析の解析用辞書である「UniDic」と「ipadic」を

それぞれ用いた結果の比較を行うと，「UniDic」の辞書を用いて形態素解析した結果の方が高い精度となった．これは「UniDic」を使うことで形態素解析の誤りが減ったため，Jaccard 係数による類似度計算がより正しくできたことにより，グラフ構造における辺の重みを改善することが可能となり精度が向上したと考えられる．

類似度計算を変更して比較した結果では，Jaccard 係数の方が高い精度となった．そのため，提案手法による学習法では cosine 類似度に比べて Jaccard 係数の方がグラフ構造の作成に適していると考えられる．

藤田らの手法を用いた文献[8]における実験結果と比較した場合，提案手法は語義曖昧性解消精度を超えることができなかった．本研究では，単語の出現形を使用していたが，藤田らの手法では素性として単語の出現形のみではなく単語の基本形も使用しているが，分類語彙表 ID は使用していなかった．また，本研究では藤田らの研究で使用していたセンスバンク「檜」の言語資源を使うことができなかった. このように使用した文脈素性に違いがあったので，精度の差が出てしまったのではないかと考える．そのため文脈素性をより近づけた状態にした方が精度の差がより無くなった可能性がある．また，精度を超えることはできなかったが，半教師あり NN は既存研究のデータ追加方法に比べると，類似度を計算すれば比較的簡単にラベル無しデータを追加する事が可能なので，これからの有効性は高いと考える．

藤田らの手法と提案手法において素性を揃えての語義識別実験を行った結果を比較すると，提案手法の方が高い精度となった．この結果は素性を揃えて半教師あり学習の学習手法について比較を行った場合，提案手法で用いたグラフニューラルネットワークの方が高い精度となった．そのため，半教師あり学習手法として提案手法の方が語義曖昧性解消の効果が高いと考えられる．

7. 結論

本稿では，グラフニューラルネットワークを用いた半教師あり語義曖昧性解消手法を提案し，提案手法が語義識別精度の改善に有効であるか分析を行った．実験の結果，SVM の結果を超えることができたので, 学習データ不足の改善が示された．そのため，グラフニューラルネットワークを用いた半教師あり語義曖昧性解消手法は語義識別精度の改善に有効であることが示された．また，今回の実験方法では，形態素解析ツールに「UniDic」を使用することや類似度計算に Jaccard 係数を使用することが語義曖昧性解消精度の向上に効果的だと示された．しかし，既存研究における半教師あり学習の語義曖昧性解消精度の結果を超えることができなかった．ただし，入力データの点で提案手法は既存研究の半教師あり学習と比べると，類似度計算をする事で比較的簡単にラベル無しデータを追加する事が可能なので今後の有効性は高いと考えられる．

今後は入力するグラフ構造に言い換え技術の利用や係り受け情報の利用を行うことやラベルなしデータのフィルタリング，岩波国語辞典の例文を効果的に適用することが課題となる．

謝辞 本研究は JSPS 科研費 18K11422 の助成を受けたものです．

参考文献

[1] 西尾実，岩淵悦太郎．水谷静夫，岩波国語辞典．岩波書店．(1994).
[2] Cortes, C. and Vapnik, V.. Support-Vector Networks. Machine Learning, Vol.20, No.3, pp.273-297 (1995).
[3] 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, vol. 22, no. 5, pp.319-362 (2015).
[4] Mihalcea, R.. Co-training and Self-training for Word Sense Disambiguation. In Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004), pp. 33–40 (2004)
[5] Sasaki, M. and Shinnou, H.. Word Sense Disambiguation Based on Distance Metric Learning from Training Documents. In Proceedings of The Sixth International Conference on Advances in Semantic Processing (SEMAPRO2012), pp. 54-58 (2012).
[6] Hu, Z. and Luo, F. and Tan, Y. and Zeng, W. and Sui, Z.. WSD-GAN: Word Sense Disambiguation Using Generative Adversarial Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 no. 01, pp. 9943-9944 (2019).
[7] Kouno, K. and Shinnou, H. and Sasaki, M. and Komiya, K.. Unsupervised Domain Adaptation for Word Sense Disambiguation using Stacked Denoising Autoencoder. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC-29), pp. 224-231 (2015).
[8] 藤田早苗，Kevin Duh，藤野昭典，平博順，進藤裕之：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 (2011).
[9] Niu, Z. and Ji, D., and Tan, C.. Word sense disambiguation using label propagation based semi-supervised learning. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL05), pp. 395-402 (2005).
[10] Yang Z. and Cohen W. W. and Salakhutdinov R.. Revisiting Semi-Supervised Learning with Graph Embeddings. In Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML2016), volume 48, pp. 40-48 (2016).
[11] Bottou, L. : Large-scale machine learning with stochastic gradient descent.， In COMPSTAT, pp. 177‑186 (2010).
[12] Okumura, M., Shirai, K., Komiya, K., Yokono, H. : Semeval-2010 task: Japanese WSD.， In: Proceedings of the SemEval-2010, ACL 2010, pp. 69‑74 (2010).