教師あり学習を用いた語義曖昧性解消における有効な情報の検討

太田 剛貴　　指導教員：山村 毅

1 はじめに

語義曖昧性解消とは、多義語が存在する文において、その多義語が文中でどういった意味で使われているかをコンピュータで識別するタスクである。機械翻訳やテキスト解析などにおいては，意味を決定する必要があることから，語義曖昧性解消は重要なタスクとなっている．

この語義曖昧性解消は，単語の意味を識別する分類問題なので，教師あり学習で解決できる．教師あり学習を用いた語義曖昧性解消は様々な研究がされているが，共通して問題となるのは「どのような情報をどのように表現し用いるか」である．

太田ら [1] は，EDR 電子化辞書[*1]の日本語コーパスにて用いられている「概念識別子」と， Mikolov ら [2] が提唱した分散表現獲得手法「Word2Vec」を用いた日本語語義曖昧性解消手法を提案している．概念識別子とは，単語の語義を 6 桁の 16 進数で表したものである．太田らはこの概念識別子を分散表現に変換することでパターン認識問題に帰着し，語義予測を試みている．

本研究では，先行研究における課題点であった，「語義予測手法」をニューラルネットワークを用いることで改良し，精度向上を目指す．

2 提案手法

先行研究では，語義予測手法に Word2Vec の CBOW モデルの出力を使用していた．しかし，CBOW モデルは分散表現獲得に適したモデルであり，語義予測に適したモデルではないため，精度は高くなかった．

そこで本研究では，モデルの出力を直接用いるのではなく，これをニューラルネットワークに通すことでこの問題に対処する．

また，概念識別子の分散表現以外の情報も伏せて利用することで，精度向上を目指す．以下，実験 1 では先行研究と語義予測手法を比較し，「概念識別子の分散表現」を用いることの有効性を検証し，実験 2 では実験 1 で用いた「概念識別子の分散表現」を別の情報と組み合わせて利用することの有効性を検証する．

3 評価実験

3.1 対象単語と対象データ

実験用のコーパスには EDR 電子化辞書の日本語コーパスを用いた．対象単語は「問題」,「前」,「場合」,「開発」,「情報」,「時間」の 6 単語 14 語義とし，日本語コーパスに与えられている概念識別子をその単語の正解の語義とする．対象データは，EDR 電子化辞書の日本語コーパスのうち，対象単語が含まれる文章 14,819 文とし，学習：検証：テストを 8：1：1 と分割して用いる．

3.2 実験 1

語義予測手法について，先行研究と比較することで「概念識別子の分散表現」を用いることの有効性を検証する．

3.2.1 概念識別子の分散表現獲得

Word2Vec の CBOW モデルを用いて分散表現を獲得する．学習用コーパスとしては日本語コーパスのうち，検証・テストデータでない文章に含まれる自立語のみを概念識別子に置き換えて用いる．また，パラメータは次元数 200，学習の窓幅 2，エポック数 10 とした．

3.2.2 分類器の学習

学習には 2 層の単純なニューラルネットワークを用いた．ニューラルネットワークの構造を図 1 に示す．

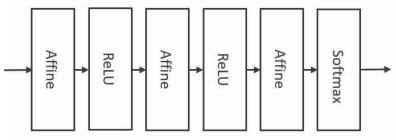

図 1　ニューラルネットワークの構造

入力層は前後 2 単語の概念識別子の分散表現の次元数 (800) に対象単語情報を One-hot ベクトルで表したもの (6) を足した 806 ユニットを持ち，中間層は 100 ユニット，出力層は語義数の 14 ユニットである．学習用データを 300 エポックで止めて，その時に得られるニューラルネットワークを分類器とした．また，損失関数は交差エントロピー誤差を使用し，最適化関数は Adam を，バッチサイズは学習データの 1/8 である 1482 を使用することとした．

3.2.3 実験 1 結果

結果は表 1 のとおりである．

表 1　実験 1 結果

ランダム	最頻出語義 (MFS)	従来手法	提案手法
42.92	75.44	72.06	79.22

従来手法と比較すると，7.16 ポイント上昇した．語義曖昧性解消に有効であるとされる MFS と比較しても提案手法が上回っており，概念識別子の分散表現を利用することの有効性を示せた．また，マクネマー検定を行うと有意水準 1% で統計的に有意であった．

3.3 実験 2

実験 1 で用いた概念識別子の分散表現を，別の情報と組み合わせて利用することの有効性を検証する．別の情報として，対象単語とその前後 2 単語の表層形を「BoW」,「Word2Vec の分散表現」という 2 つで表現したものを利用する．なお，対象データとニューラルネットワークの構造は実験 1 と同様のものを使用する．

3.3.1 BoW

BoW（Bag of words）とは，文書中の単語の出現数を数え上げるカウントベースの手法のことである．対象データに含まれる自立語は 20,524 単語であった．単語にインデックスを割り当て，素性として用いる単語の出現回数を対応する単語のインデックスに割り当てることで，それを素性の特長ベクトルとする．したがって，入力層は対象単語とその前後 2 単語を BoW で表現し

[*1] https://www2.nict.go.jp/ipp/EDR/JPN/J_indexTop.html

た次元数 20,524 ユニットであり，概念識別子の分散表現を組み合わせる場合は，次元数 800 を足した 28,524 ユニットである．

3.3.2 Word2Vec の分散表現

表層形の分散表現獲得には，Word2Vec の CBOW モデルを用いる．学習用コーパスとしては日本語コーパスのうち，検証・テストデータでない文章の自立語のみを用いる．また，パラメータは次元数 200，学習の窓幅 2，エポック数 10 とした．入力層は対象単語とその前後 2 単語の分散表現を連結した次元数 1,000 ユニットであり，概念識別子の分散表現を組み合わせる場合は，次元数 800 を足した 1,800 ユニットである．

3.3.3 実験 2 結果

結果は表 2 のとおりである．なお，表では概念識別子の分散表現を CIE（Concept Identifier Embedding）と記載している．

表 2　実験 2 結果

BoW	BoW + CIE	Word2Vec	Word2Vec + CIE
79.96	79.35	82.73	83.67

まず表層形の表現方法について比較すると，表現方法として Word2Vec を利用する方がよいと分かった．

また，マクネマー検定を行うと，BoW の場合は $p = 0.48$，Word2Vec の場合は $p = 0.17$ であり，どちらの場合でも有意水準 1 %で有意差がないという結果になった．このことから，概念識別子の情報量は単語の表層形と大きな差がないことが分かり，概念識別子の精度に問題があると考えた．

4 考察

本研究では，概念識別子の分散表現を用いているため，その概念識別子の精度が非常に重要である．そのため，本節ではこの概念識別子について考察する．

4.1 概念識別子が与えられていない単語について

日本語コーパスには構成要素情報として，「概念選択」が与えられている．この概念選択には，1. 概念識別子，2. 補足付き概念選択，3. 複合語形態素番号，4. 記述なし の 4 つの場合があり，本研究ではこのうちの概念識別子を利用していた．日本語コーパスに現れる単語に与えられている，概念選択の割合を表 3 に示す．

表 3　日本語コーパスに現れる単語の概念選択の割合

概念選択	単語数	割合
概念識別子	4,519,834	89.1%
補足付き概念選択	350,546	6.9%
複合語形態素番号	53,998	1.1%
記述なし	149,821	3.0%

対象データとしては，複合語形態素番号が含まれている文は用いておらず，単語に概念識別子が与えられていない場合はその単語を素性から省いていた．

したがって，概念識別子が与えられていない単語に対して，概念識別子のようなラベルを与えることができれば，より質の高い概念識別子の分散表現を作成できると考えられる．

4.2 文章による概念選択の違い

同じ意味の単語には同じ概念選択が与えられていることが理想的であるが，文章によって概念選択が異なっている場合があった．3 つの文章に含まれる「趙紫陽」という単語に与えられた概念選択の例を表 4 に示す．

表 4　「趙紫陽」という単語の概念選択

文章	概念選択
1	”＝Ｚ　趙紫陽という人”
2	”＝Ｚ　趙紫陽という中国の政治家
3	”＜ｃ＃ｈｕｍａｎ”

また，概念識別子ではなく補足付き概念説明が与えられているものは，単語辞書に適切な概念が存在していない場合とされていたが，「趙紫陽」という単語は単語辞書に存在し，概念識別子が与えられていた．そのため，単語辞書と照らし合わせながら，概念選択を修正してから用いる必要があると考える．

4.3 概念識別子の曖昧性

EDR 日本語単語辞書に載っている「問題」という単語に与えれている概念識別子をすべて取り出し，概念識別子について考察する．「問題」に与えられている概念識別子とその概念説明を表 5 に示す．

表 5　「問題」の概念識別子とその概念説明

概念識別子	概念説明
10bd37	解決を求めて取り上げるべき事柄
10bd3a	難点
3ce84f	解決するのがむずしい問題
3cec3d	当然しなければならないこと
3cfc60	面倒で嫌なもの
2025ca	試験問題
2025cb	世間の注目を浴びている事柄
3d00ad	物事の一番大事なところ

概念説明を比較すると，それぞれの差が曖昧であるように思われる．そのため，より質の高い概念識別子の分散表現を作成するためには，概念識別子を修正してから用いる必要があると考える．

5 終わりに

本研究では，日本語語義曖昧性解消のタスクに対し，概念識別子を利用することの有効性について検討した．ニューラルネットワークを用いて語義予測した結果，先行研究を 7.16 ポイント上回る，79.22% の正解率を記録し，概念識別子を利用することの有効性を示した．

しかしながら，概念識別子の分散表現を別の情報と組み合わせて利用しても，大きな効果がなかった．

今後の課題としては，概念識別子を修正してから用いることや対象単語・対象データの増加，文脈を考慮した概念識別子の分散表現の利用について考えていきたい．

参考文献

[1] 太田剛貴, 山村毅: ”概念識別子と文の依存構造を用いた語義曖昧性解消”, 人工知能学会研究会資料 言語・音声理解と対話処理研究会, SIG-SLUD-092-07, オンライン開催, 2021

[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean : “Efficient Estimation of Word Representations in Vector Space”, Proceedings of ICLRWorkshop 2013, pp.1-12, 2013.