# Word2Vec を用いた非線形日本語語義曖昧性解消

情報科学科　加藤 敦也　　指導教員：山村 毅

## 1　はじめに

語義曖昧性解消とは，文に現れたある単語が，その文脈においてどのような語義を持っているかを判断するための，自然言語処理のタスクの一つである．現在注目されている，機械翻訳やテキスト解析などにおいて，語義曖昧性解消は重要なタスクとなっている．近年，自然言語処理の分野では単語の意味表現をベクトルで表す『単語の分散表現』を用いた研究が進められている．その 1 つとして，加藤ら [1] は，EDR 電子化辞書[*1]の日本語単語辞書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．この研究では，49.8% の曖昧性解消の精度を記録しており，手法の有効性を示している．

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』を改良し，更なる曖昧性解消の精度向上を目指すこととする．

## 2　提案手法

### 2.1　前処理と学習モデル作成

本研究では EDR 電子化辞書の日本語コーパスを実験用データとして用いる．このコーパスに含まれる文は，すべて分かち書きされており，各単語に対してその文における単語の語義を示す『概念識別子』が 1 つ割り振られている．概念識別子とは“3ce84f”のような，単語の語義を一意に示すための 16 進数のコードのことである．そこで，前処理として学習データで使う文に関しては全ての単語を概念識別子に置き換える．そして，この学習データに Word2Vec の CBOW モデルを用いて，概念識別子の学習モデル，すなわちベクトル表現を獲得する．

### 2.2　学習モデルを用いた語義予測手法

前節にて作成したモデルとテストデータを用いて，語義の予測を行う．テストデータの文は，曖昧性解消の対象単語以外の全ての単語を概念識別子のベクトルに置き換えた文を使用する．このテストデータの文に対して，学習モデルを用いて式 (1) の計算を行う．

$$C_{Predict} = \underset{C_{Target} \in C_1, C_2, \ldots, C_k}{\arg\max} P(C_{Target} | C_{Arround}) \quad (1)$$

ここで $C_{Predict}$ は対象単語の予測概念識別子のベクトル，$C_{Target}$ は対象単語が持つ概念識別子 (EDR 電子化辞書の日本語単語辞書に記述されている) のベクトル，$C_{Arround}$ は対象単語の前後 N 単語の概念識別子のベクトルから計算される平均ベクトルである．式 (1) の計算には，Word2Vec の CBOW モデルの出力を利用している．学習する際に用いた CBOW モデルに則り，『前後 N 単語の概念識別子が入力されたときにどの概念識別子が対象単語の語義としてふさわしいか』を計算すれば語義を予測できるのではないか，と考えた．求めた $C_{Predict}$ が，対象単語の持つ真の概念識別子と等しい場合，正解とする．

## 3　実験方法

Word2Vec のパラメータは次元数 200，学習の窓幅 5，学習の反復回数 10 とし，3 回未満しか登場しない概念識別子については学習しないこととした．曖昧性解消の対象単語は，文の中心に最も近い多義語の名詞とした．対象とした単語は 167,661 個であり，精度の検証方法としては 10 分割交差検証を用いた．

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

## 4　結果と考察

### 4.1　実験結果

3 章にて述べた方法を用いて，提案手法の精度評価を行った．結果を表 4.1 に示す．比較のために，本研究のデータに対して，従来手法 [1] を適用した結果も示す．

| 手法 | 正解率 [%] |
|---|---|
| 従来手法 | 24.6 |
| 提案手法 | 59.5 |

表 1　従来手法と提案手法における実験の正解率

曖昧性解消の対象単語が持つ概念識別子の平均数は 3.92 個であるため，無作為に概念識別子を選んだ時の正解率は 25.5% である．本研究ではそれを明らかに上回る結果となっている．

### 4.2　考察

まず，従来手法と提案手法の結果の比較から行う．提案手法の正解率が，従来手法の正解率を 34.9 ポイント上回っていることから，本手法の方が日本語語義曖昧性解消に有効であると言える．従来手法による精度が文献 [1] で行った実験のときよりも大幅に落ちたのは，対象単語の品詞を指定しておらず，付属語といった意味をあまり持たない単語まで曖昧性解消の対象単語に選んでしまっていたからであると考える．

また，あるテストデータの対象単語が「モスクワ会談」の『会談』であるとき，語義の候補は“interview with someone”，“conference”を示す概念識別子と，どちらも似た意味であるが，提案手法ではこれに対し，正しい概念識別子を予測できたが，従来手法では誤った予測をした．これはベクトルの線形的な近さを予測手法として使ったのではなく，非線形な Word2Vec の出力を用いる予測手法を使ったからであると考えている．

## 5　まとめ

本研究では Word2Vec の学習モデルを用いた非線形的な語義予測手法を提案して，従来手法よりも高い精度の日本語語義曖昧性解消を行うことができることを明らかにした．しかし，その精度は約 60% と決して高いとは言えない．新たな曖昧性解消の素性の導入や，コーパスを追加することで更なる精度の向上ができるのではないかと見込んでいる．

## 参考文献

[1] 加藤 敦也，内藤 弘章，山村 毅：”概念識別子の分散表現を活用した語義曖昧性解消法”，電気・電子・情報関係学会東海支部連合大会講演論文集，J1-3，大同大学，2019.

[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.