4W-01

機械学習を用いた単語の意味の分類
Semantic Classification of Words Using Machine Learning

竹嶋翔矢† 篠山学† 松本和幸‡
香川高等専門学校† 徳島大学‡

1. はじめに

近年，自然言語処理や画像処理に深層学習が盛んに利用されている．深層学習が使用されている有名なシステムには，Siri，Alexa，Google アシスタントなどのスマートスピーカーやソフトバンクが開発した Pepper[4]などがある．それらのシステムは，深層学習の利用により実用化されたものである．

特に最近，自然言語処理タスクで精度が高いとされているTransformersモデルは，マスク言語モデルと呼ばれるモデルであり，大量のテキストデータによる単語の穴埋め問題を事前学習タスクに用いている． また，BERTと呼ばれる，Transformersモデルをベースとして，自然言語処理において文脈を考慮したモデルがある．本研究で用いるGiNZAのELECTRAモデルは，BERTを改良したELECTRAをベースとしてGiNZA[1]に組み込んだものである．

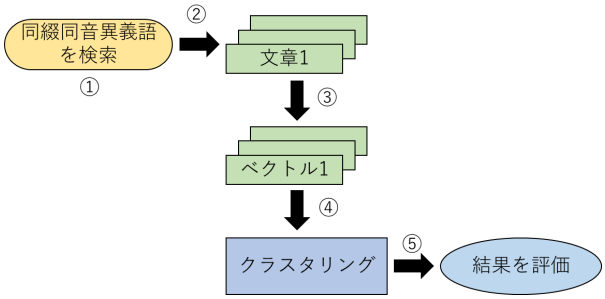
深層学習により自然言語処理の多くの課題が解決しつつあるといえる．しかし，まだ解決できない問題も多い．現在，自然言語処理では，同じ綴り，同じ発音であっても文脈や前後の単語によって意味が異なる単語（以後，同綴同音異義語と呼ぶ）の解釈が難しい．

本研究では， 同綴同音異義語の意味の分類を目指す．例えば「頭」には「体の一部」と「物事のはじめ」の意味があり，文脈によって意味が異なる．このような語を意味ごとに分類する手法を提案する．また，GiNZA の通常モデルと文脈を考慮した ELECTRA モデルとを比較する．

2. 提案手法

本研究の提案手法を図 1 に示す．

① 中納言[2]と呼ばれる国立国語研究所が開発した日本語コーパスを検索するための Web アプリケーションを用いて同綴同音異義語を検索する
② 検索結果に表示された文章から同綴同音異義語を含む文章を収集する



1 本研究の提案手法

③ 収集した文章を GiNZA を使用してベクトルに変換する
④ ③で作成されたベクトルを用いて，教師あり学習のランダムフォレスト法により文章の分類を行う
⑤ ④で得られた分類結果の精度を混同行列，または正解率を用いて評価する

3. 文章の収集方法

本研究で対象とする同綴同音異義語は「頭」とした．「頭」には，「体の一部」の意味で使用されている場合と「物事のはじめ」の意味で使用されている場合がある．それぞれの意味で「頭」が使用されている文章を収集する．ここで「物事のはじめ」の意味で「頭」が使用されている文章は文章数が少ないことから「最初」という単語を「頭」に置換することで文章を収集した．

文章の長さは，同綴同音異義語の前後 20 文字とした．収集した文章数は，意味ごとに 500 文，合計 1000 文とした．

4. 文章のベクトル化

文章のベクトル化には，spaCy[3]をベースにした GiNZA という日本語自然言語処理ライブラリを用いる．GiNZA は，形態素解析処理に SudachiPy を使用しており，単語のベクトル表現に chiVe（Sudaci Vector）と呼ばれるスキップグラムアルゴリズムに基づいて，word2vec を使用してトレーニングをした単語のベクトルがあらかじめ用意されている．単語のベクトルの次元数は 100 次元である．また，文章のベクトルの次元数も 100 次元で，各単語の 100 次元のベクトルを平均化したものとなっている．

Semantic Classification of Words Using Machine Learning
†Shoya Takeshima, †Manabu Sasayama, National Institute Of Technology. Kagawa College
‡Kazuyuki Matsumoto, Tokushima University

5. 特徴量の追加

収集した文以外の文章から同綴同音異義語の意味を判別できる特徴量を3つ追加する．以下に示す．

特徴量1 ：「頭」の直後に助詞の「～の」「～に」「～から」がつく
特徴量2 ：「頭」の直後に助詞の「～を」がつく
特徴量3 ：「頭」と係り受け関係のある単語の品詞

6. 評価実験

提案手法の有効性を確認するために，人の体を表す同綴同音異義語を対象とし，同綴同音異義語の意味を教師あり学習により分類するモデルの作成を目指す．本研究では，「頭」を対象として収集した文章1000文を用いて，評価実験を行う．1000文をGiNZAにより100次元のベクトルに変換し，ランダムフォレスト法を用いて分類する．文章のデータを訓練用データ 70%，テスト用データ 30%として分類する．評価実験の結果を混同行列と正解率を求めることで評価する．正解率の計算式を式1に示す．

$$\text{正解率} = \frac{\text{正しく分類された文章数}}{\text{すべての文章数}} \quad (1)$$

7. 評価実験の結果

評価実験の混同行列を表1に示す．

表1　評価実験の混同行列

	体の一部	物事のはじめ
体の一部	**126**	**31**
物事のはじめ	**37**	**106**

評価実験の結果より，正解率は 0.773 となった．

8. 比較実験

GiNZAのELECTRAモデルを用いた実験の提案手法を図2に示す．

比較実験では，評価実験で収集した同綴同音異義語の前後20文字の文章から意味ごとに10文ずつ，合計 20 文を使用する．「体の一部」の意味で「頭」が使用される文章を‘文章 1~10’，「物事のはじめ」の意味で使用される文を‘文章 11~20’とし，文章1~20中の「頭」の単語ベクトルを‘vec1~vec20’とする．

初めに，vec1~10 と vec11~20 の単語ベクトルの cos 類似度とその平均を求める．vec5 と vec11~20 の cos 類似度の平均を表 2，vec7 と vec11~20 の cos 類似度の平均を表 3 に示す．

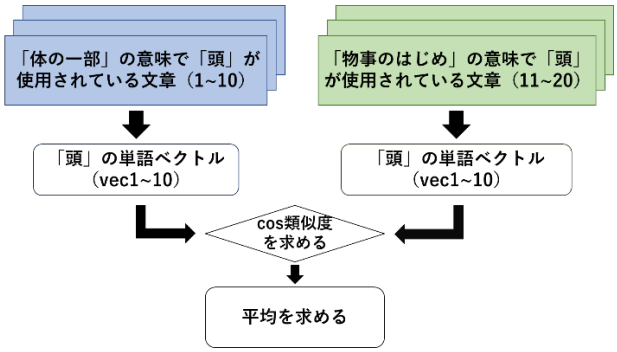


図2　比較実験の提案手法

表2　vec5 と vec11~20 の cos 類似度の平均

	vec5
cos 類似度の平均	**-0.147022**

表3　vec7 と vec11~20 の cos 類似度の平均

	vec7
cos 類似度の平均	**0.854325**

表2より，vec5 と vec11~20 の cos 類似度の平均はマイナス値となっている．このことから「体の一部」の意味を持つ「頭」と「物事のはじめ」の意味を持つ「頭」の単語ベクトルが正反対に位置し，意味の判別ができていると考えられる．しかし，表3を見ると，vec7 と vec11~20 の cos 類似度の平均を見ると，「体の一部」の意味を持つ「頭」と「物事のはじめ」の意味を持つ「頭」の単語ベクトルが近い位置に存在し，意味の判別ができていないことがわかる．vec1~6 と vec8~10 も vec7 の結果に近い結果であった．文脈を考慮したモデルでも単語の意味を分類できていないことがわかる．

9. おわりに

本研究では，同綴同音異義語を判別できるモデルの作成を目的とし，「頭」を対象として評価実験を行った．結果より，「頭」に関しては，正解率が約8割と高い結果が得られた．

謝辞

本研究は JSPS 科研費 19K12174 の助成を受けたものです．

参考文献

[1] GiNZA：https://megagonlabs.github.io/ginza/
[2] 中納言：https://ccd.ninjal.ac.jp/bccwj/
[3] spaCy：https://spacy.io/