語の用例の外れ度合いに基づく多義語の難しさ評価と視覚的な学習支援
Word Usage Difficulty based on Outlier Detection

江原 遥
Yo EHARA
東京学芸大学
Faculty of Education, Tokyo Gakugei University
Email: ehara@u-gakugei.ac.jp

あらまし： 外国語語彙学習支援において複数の意味を持つ多義語は難しさを定めにくい．直感的には語義ごとに難しさが異なるが，語義数や分類に言語学的な統一見解がない事も多い．そこで本稿では，語義の定義の問題を迂回し，語の用例の意味的外れ度合いを用例の難しさとする手法を提案する．多義語の各語義の知識を問う語彙テストを作成した．その結果による評価実験では，提案法は従来法より統計的有意に高精度で被験者反応を予測した．

1 はじめに

本稿の内容は AIED 2022 full paper に採択されたものである [5]．外国語学習において，語彙学習は学習者が学ぶのに必要な時間が長いうえ，読解力をはじめとする全般的な語学力と相関が高いため，特に支援を要する．語彙学習の支援においては，学習者が適切な語の使い方を学べるよう，各単語の主要な使い方（用例）を学習者に提示したいニーズがある．母語話者の作文や発話を集めた大規模コーパスは，均衡コーパスなどの形で多くの言語で容易に入手可能であるので，こうしたコーパス中の，ある単語の出現のうち，どの出現が学習者が覚えるべき主要な用例に相当し，どの出現が例外的であるのかがわかれば，学習者にとって有用と思われる．

この時，単にコーパス中の当該単語の出現箇所を羅列するのではなく，多義語については語義を考慮し，類似した語義を持つ出現をまとめて提示してくれる機能や，覚えるべき主要な語義の出現と，例外的な語義の出現を分けて提示してくれる機があると，より語彙学習に望ましい．しかし，このように．語の出現ごとに語義を付与したり，覚えるべきかどうかを判定する作業を，人手で行うことは非現実的である．

語義については，近年，文脈を考慮して単語の各出現（用例）ごとに，異なる埋め込みベクトル表現を求める「文脈化単語埋め込み」の手法が，主に自然言語理解のタスクにおいて有力である [3]．文脈化単語埋め込みベクトルは出現ごとの意味的情報を含んでいるため，前述の望ましい機能を人手のコストなしに実現可能であるように思われる．教育現場でも理解されるためには，まず，ベクトルを人間に可視化する機能が必要であろう．次に，語彙学習のためには主要な語義と例外的な語義を分けて提示してくれる仕組みが欲しい．数百次元ほどある文脈化単語埋め込みベクトルの集合の外れ値の検出に有効と思われる手法として，「教師なし深層異常検知」[6] が挙げられる．この手法では，深層学習によってデータを低次元に圧縮し，クラスタリングを行いながら，どのクラスタからも離れている点を外れ値（異常）として検出する．

そこで，本研究では，文脈化単語埋め込み [3] と教師なし深層異常検知 [6] に基づき，人手のアノテーション情報なしで前述の機能を実現することで，語の多義性・主要性を学習者に提示する手法を提案する．そして，実際に多義語に関する語彙テスト結果のデータセットを作成し，これを用いて提案手法を評価する．

2 深層異常検知

深層異常検知の近年の代表的な手法として，DAGMM[6] が挙げられる．DAGMM は，クラスタリング手法として有名な混合ガウスモデル（Gaussian Mixture Model, GMM）を深層化し，異常検知の機能を持たせた手法である．高次元ベクトルを次元圧縮し，低次元表現で GMM に基づくクラスタリングをした上，直感的には各クラスタ中心からの距離の和として理解

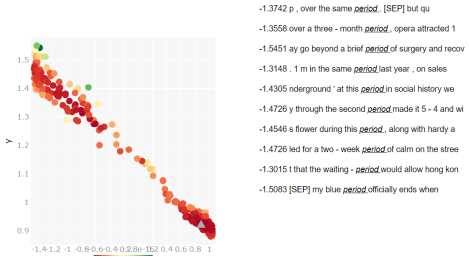

図 1: “period” の主要な用例．丸い各点は，“period” の各用例（コーパス中の各出現）に対応する．各点の色は例外的である度合い（エネルギー値）を表し，緑色ほど例外的，赤色ほど主要と判定されている．右下の赤い点が多く集まる部分にある，灰色の▲が基準点であり，基準点からの距離が近い点 10 点に対応する用例が，テキストの形で右側に示されている．テキストの前の数値は，実際の各用例のエネルギー値である．本稿の文例は，全て BNC[2] から取得した．

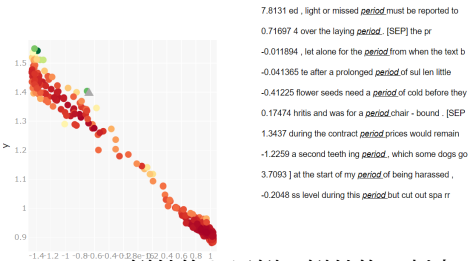

図 2: “period” の例外的な用例．例外的と判定された緑色の点に合わせて基準点を設定し，この緑色の点に対応するテキストが，右側の一番上に表示されている．

できる「エネルギー値」を計算し，どのクラスタ中心からも遠い点を異常として検知する．

DAGMM は，入力ベクトル $\vec{x}$ を自己符号化器を用いて低次元表現 $\vec{z}$ に変換し，$\vec{z}$ から $\vec{x}$ を再構成する深層学習モデルである．再構成したベクトルを $\vec{x}' = g(\vec{z}_c; \theta_d)$ とし，低次元表現を $\vec{z}_c = h(\vec{x}; \theta_e)$ とする．再構成したベクトルと元の入力の近さを測る関数を $\vec{z}_r = f(\vec{x}, \vec{x}')$ とする．ここで，この近さとしては複数の関数が利用できる．DAGMM は，低次元表現と再構成の誤差をつなげた $\vec{z} = [\vec{z}_c, \vec{z}_r]$ を最終的な用例の潜在表現として利用する．最終的に語の用例のエネルギー値 $E(\vec{z})$ を計算し，これが高いほど外れ度合いが高いと判定される．

3 視覚的学習支援・頻度修正実験

イギリス英語の均衡コーパスとして，代表的な British National Corpus（BNC）[2] のうち，10 万文に対して BERT[3] を適用し，最も上位の層（出力に近い層）から文脈化単語埋め込みベクトルを得た．BERT モデルとし

ては，**bert-base-uncased** を用いた[1]．文脈化単語埋め込みベクトルの次元数は 768 である．入力された単語に対して，対象データ中の全単語の出現と，各出現に対応する文脈化単語埋め込みを取得できるようにした．

実装は，第三者によって公開されている DAGMM の PyTorch 実装をもとに行った[2]．訓練のハイパーパラメタは，次元数と DAGMM のクラスタ数を 3 と設定設定した…以外は，この実装と同一である．

“period” という語を例に議論する．紙面での見やすさを考慮して．DAGMM による用例の潜在表現 $\vec{z}$ の 3 次元表現の最初の 2 次元分を用い，図 1 と図 2 に “period” の語の用例の可視化例を示した．対象の 10 万文中，“period” は 376 回出現した．各点が “period” の各出現の文脈化単語ベクトルを 2 次元座標上で表現したものであり，各出現に対応している．各点の色はエネルギーの値を表す．この値は高いほど例外的，すなわち，緑色ほど例外的と判定されている．逆に赤いところほど主要な用例と判定されており，直感的にはヒートマップと同等に解釈できる．

横軸・縦軸は，それぞれ，DAGMM の潜在空間表現 $\vec{z}$ の第 1 次元，第 2 次元である．灰色の三角形の点は基準点であり，この点に図上で最も近い順に，これに対応する 10 点に対応するテキスト 10 件が用例として右側に提示される．用例の左側にあるのは，実際に計算されたエネルギー値である．基準点はマウスでドラッグして動かせるようになっており，学習者は興味のある点の近くに基準点を移動させることによって，どのような用例があるのかを把握できる．

まず，図 1 を見ると，2 つのクラスタに分かれていることがわかる．この可視化が，各用例の語義を反映していれば，学習者にとって望ましい機能のためには有用であろう．しかし，元の高次元ベクトルを 2 次元で表現することは難しく，各クラスタが語義を反映していないこともある．そのような結果であっても，学習者にとっては，学習の優先度が高い用例が示されていれば，学習優先度が分かるという点では有用であろう．図 1 では，各点の属しているクラスタに関わらず，クラスタの中心部分が赤く，クラスタの端の部分が外れ値として判定されていることがわかる．図 1 には，基準点をクラスタの中心部分に置いた場合の例を示す．基準点の周りの，「期間」という広く知られた意味の “period” の用例が右側に並べられている．このように，深層異常検知によって，外れ度合いが低い語を，語の主要な用例として提示する事が可能であることが示されている．図 2 には，緑色の例外的な用例の例を示す．右側には “light or missed period” という用例が出ている．“period” には，「期間」という意味の他に，「生理」という意味があり，これは「軽い，または来なかった生理」と訳されるものである．この意味での “period” は，少なくとも “period” の主要な用例ではなく，例外的な用例と判定されていることがわかる．また，この例外的と判定された用例でも，“period” は名詞として使われており，固有表現の一部などでもない．従って，この用例は，品詞推定や固有表現抽出を用いて捉えることは難しい．

最後に，各単語の外れ度合いの閾値をパラメタとして，DAGMM との同時学習により閾値未満の出現のみを単語頻度とみなし頻度修正を行いながら学習する多層ロジスティック回帰を実装し，精度評価を行った．100 語種について 100 人をテストした単語テストデータ [4] を用い，23 語 ×100 人，計 2,300 件を訓練，10 語 ×100 人，計 1,000 件をテストに用い，学習者の単語テストの正答/誤答の予測精度を用いて評価した．BNC 中の単語頻度をそのまま特徴量に用いた場合と，外れ度合いを用いた頻度修正を行った場合では，どちらも精度は 0.75 であった．従って，提案手法は既存手法と同等の精度を達成しながら，図 1 や図 2 に示す詳細な分析が可能であることが示された．

[1] https://github.com/huggingface/transformers
[2] https://github.com/danieltan07/dagmm

表 1: 被験者反応の予測精度.

	典型的な語義	例外的な語義
Size-based	62.8%	44.4%
LR	63.1%	47.9%
提案法	63.7% (*)	48.5% (*)

4　外れ度合いによる難易度の評価実験

外れ度合いにより頻度を修正する形ではなく，直接外れ度合い $E(\vec{z})$ を特徴量に用いて難易度として用いた場合，どのような結果が得られるだろうか？この問いに答えるため，各単語の例外的な意味と典型的な意味の両方について受験者が回答する特殊なデータセットを作成して実験を行った．このデータセット（`http://yoehara.com/` にて公開予定）には，58 個の典型的な語彙の質問と 12 個の単語の例外的な意味と典型的な意味の質問のペアが含まれている．質問は複数の英語母語話者に確認済みである．58 問の典型語彙問題と 12 組の例外・典型語法問題に対する 235 名の被験者の回答をそれぞれ学習データとテストデータとして使用した．

“Size-based” は語彙量に基づく手法であり，推定した語彙量に基づき，均衡コーパス上での単語頻度順位の昇順で推定語彙量までは全部知っている，それ以降は全部知らないと推定する手法である [1]．この方法では，まず学習者の語彙量を推計し，推計語彙量よりも高頻度語は全て知っている（またその逆も真）という手法である．**LR** はロジスティック回帰を用いた手法である．特徴量には均衡コーパスとして代表的な BNC，Corpus of Contemporary American English (CoCA) コーパス中の単語頻度，並びに学習者 ID を特徴量に用いた．学習者 ID によって個人化識別に対応させている．提案法は，コーパス特徴量に加え，質問文内の対象用例の $E(\vec{z})$ を特徴量に追加している．

表 1 に示された結果は，提案法が典型的な語義，例外的な語義の双方において，従来法を予測精度で上回ることを示す．これは，文例の外れ度合いの特徴量が有効に働いているためと思われる．(*) は，提案法と **LR** の差が統計的に有意であることを示す（Wilcoxon, $p < 0.01$）．

5　おわりに

本稿では，厳密に分類しにくい語義という学習項目について，その難しさを深層異常検知を用いて求める手法を提案し，実データで提案法の有効性を評価し，外れ値を視覚化する手法も提示した．本稿では語彙学習支援に限定したが，原理的には厳密に分類しにくい学習項目一般に本研究の手法は適用可能であるので，「外れ値は難しい」という発想で他分野でも追加の実証・応用につなげることが今後の課題として期待される．

謝辞

本研究は，科学技術振興機構 ACT-X 研究費 (JPMJAX2006) の支援を受けた．

参考文献

(1) David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, Vol. 31, No. 7, pp. 9–13, 2007.
(2) BNC Consortium. *The British National Corpus*. 2007.
(3) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pp. 4171–4186, Minneapolis, Minnesota, June 2019.
(4) Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.
(5) Yo Ehara. An intelligent interactive support system for word usage learning in second languages. In *Proc. of AIED*, 2022.
(6) Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.