外国語語彙学習における意味的多様性の高い多義語用例の提示システム

江原　遥[1,a)]

概要：外国語学習においては，多くの語を学習する事が必要になる．外国語の語と母語ではそもそもの使い方が異なるため，単純に語とその母語での翻訳を見せるのではなく，母語話者が実際に語を用いた箇所をテキストから抜き出して「用例」として提示することが重要である．さて，学習者が初期に学ぶ語の多くが，複数の異なる意味を持つ多義語である．意味が複数あり，それぞれよく使われる場合は，多様な用例を提示してくれることが望ましい．しかし，個々の用例の意味を人手で付与することはアノテーションコストの面から非現実的である．また，教育の目的では，意味付与を自動で行い学習者に提示するアプローチも，誤りを含む可能性があるため難しい．そこで，本研究では，用例に対して，意味的多様性を考慮した「主要度」を自動付与する手法を提案する．学習者は，主要度の高い順に用例を学ぶだけで，その語の多様な意味を学ぶことができる．提案手法は教師データを必要としないため，様々なドメインや言語に適用できる利点がある．まず，事前学習済言語モデルと転移学習の考え方を用いることで，語の出現（用例）ごとの意味を捉える文脈化単語埋め込みベクトル集合を語ごとに作成する．次に，各用例の文脈化単語埋め込みベクトルが外れ値である度合い（異常度）を計算する事で，各用例がどの程度重要な用例かを計算する．実際に日本語を母語とする英語学習者に対して，英語母語話者により妥当性が確認された多義語の語彙テストを行うことで知る限り初めて第二言語学習者の多義語の語彙知識に関する公開データセットを作成した．評価の結果，有望な結果を得た．

1. はじめに

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，次のような提示を行うと，より語彙学習に有用であると予想される．

(1) 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

(2) 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能

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，アノテーションコストが高すぎ，非現実的である．

語の各出現に対して自動的に語義を付与するタスクは，古くから語義曖昧性解消（Word Sense Disambiguation, WSD）として研究されている．しかし，WSD の技術を直接利用した上記機能の実現は，次の理由により難しい．まず，多くの WSD の研究は，一部の頻出語かつ多義語である語だけを対象にしている．一方，機能 1 は学習者が調べたい全ての語に対して行える事が望ましい．全ての語を対象とした WSD である **all-words WSD** の研究は，比較的少ない [17]．次に，WSD は，正確に語義を付与することが主目的であり，WSD の技術からは，語義ごとの主要性（学習上の優先度）を付与する機能 2 は直接的には実現されない．例えば，語義が 3 種類あり，そのうち 2 つは主要だが意味が似通っており，1 つの語義はほとんど出てこない場合，学習者の観点からは，主要な 2 つの語義間の正確な分類よりも，各出現が主要な語義の出現であるのか，例外的であるのかの方に関心があると思われる．このように，語

[1] 東京学芸大学
Tokyo Gakugei University, Tokyo Japan. 184–8501, Japan.
[a)] ehara@u-gakugei.ac.jp

の各出現の主要性判定は，WSD からは直接実現される機能ではなく，WSD とは異なる仕組みが必要になろう．最後に，WSD のうち，人間のアノテータが人手で付与した語義ラベルを教師データとして用いる教師あり WSD の設定では，教師データ作成に，やはり高いアノテーションコストが必要になる問題がある．

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大きなブレイクスルーを起こしており [2], [8]，語義曖昧性解消にもすでに利用されている [16]．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，上記の 1，2 の機能を実現する上で重要であると思われる．しかし，文脈化単語埋め込みベクトルは，通常，数百次元程度の高次元ベクトルであるため，そのまま提示しても外国語学習者は理解できない．文脈化単語埋め込みベクトルを用いて，上記の 1，2 の機能を実現するためには，まず，文脈化単語埋め込みベクトルを次元圧縮し，可視化することが必要になる．次に，1 の実現のため可視化空間でのクラスタリング，2 の実現のために各ベクトルの主要度の計算の，3 種のタスクを行う必要があろう．この 3 種のタスクを同時に行う手法として，「教師なし深層異常検知」[10] が挙げられる．この手法では，深層学習によってデータを可視化次元に圧縮し，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データ点を外れ値（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 [2] と教師なし深層異常検知 [10] に基づき，人手のアノテーション情報なしで 1，2 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．また，本研究のために，著者が知る限り初めて，第二言語学習者の語義ごとの語彙知識を確認するテストを英語母語話者のチェックのもと作成し，データセットを作成した．

本稿の貢献は，提案手法の有用性を定性的に示したことである．具体的には，“period” という 1 つの主要な語義と 1 つの例外的な語義を持つ語に対して，主要性と例外性が提案手法により認識されることを示した．また，“figure” という複数の主要な語義を持つ語に対して提案手法を適用することで，提案手法により，WSD などと異なり，複数の主要な語義が適切に「主要」と判定されていることを示したことである．また，“seven-figure” と “six-figure” という，使用頻度に差があると思われる表現が，それぞれ，前者が例外，後者が主要，と，直感に沿う形で主要性が認識されていることも示した．

2. 関連研究

2.1 語彙学習支援システム

語彙学習支援システムについては，[15] に詳細な関連研究がまとめられている．また，外国語の語彙学習のための良い教育法については，応用言語学分野で研究されている．例えば，類似した複数の語義の背後に核となる「コアミーニング」があることを仮定し，これを核に，派生する意味を教える教育法などが提案されている．こうした応用言語学分野の知見については，[11] が詳しい．

2.2 複雑単語推定と既習語予測問題

テキスト中で語学学習者にとって難しい語を探すタスクは，自然言語処理分野では Complex Word Identification (CWI，複雑単語推定) と呼ばれる [7], [9]．複雑単語推定は，必ずしも語学学習者が語を知っているかどうかには関わらず，母語話者や語学学習者が，他の語学学習者にとって難しいと思われる語を発見するタスクである．これは，複雑単語推定の目的がテキスト簡単化であるため，語を知っているかどうかは副次的な問題であるためと推測される．複雑単語推定と類似しているが，異なる設定のタスクとして，学習者が知っている単語を探す既習語予測問題が挙げられる．これは，覚えるべき語の学習者への提示を目的としている．既習語予測問題に関しては，[13] の報告がある．この方法では，やはり，既習語予測問題をニューラルなモデルとして定式化している．

2.3 用例情報提示

語彙学習のために文脈化単語埋め込みに基づき，用例を提示する研究は，[4], [12], [13], [14] が挙げられる．このうち，[4], [12] では，単純に主成分分析（Principal Component Analysis, PCA）を用いて 2 次元に可視化していたが，クラスタリングは行われておらず，どの単語がどの程度重要かも示されていない．

学習者の単語テストのデータから，文脈化単語埋め込みの可視化自体を学習する方法はあるが [5], [13]，単純に線形写像が用いられており，本稿で述べる非線形な可視化は行われていない．[14] では，深層異常検知と単語テストデータを組み合わせて可視化を学習するモデルが提案されているが，実験結果の可視化例は言及されていない．

3. 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[10] 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．語義曖昧性解消に関連して，文脈化単語埋め込み表現をクラスタリングして，各クラスタを語義とみなしてまとめる手法

が提案されている [16]．GMM はクラスタリングの代表的な手法であるため語義曖昧性解消の既存研究との親和性・解釈性を考慮して，DAGMM を選択した．DAGMM は自然言語処理では応用例は少なく，知る限り他に固有表現抽出での応用例があるのみである [6]．

DAGMM は，入力ベクトル $\boldsymbol{x}$ をオートエンコーダを用いて低次元表現 $\boldsymbol{z}$ に変換し，$\boldsymbol{z}$ から $\boldsymbol{x}$ 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 $\boldsymbol{x}' = g(\boldsymbol{z}_c; \theta_d)$ とし，低次元表現を $\boldsymbol{z}_c = h(\boldsymbol{x}; \theta_e)$ 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 $\boldsymbol{z}_r = f(\boldsymbol{x}, \boldsymbol{x}')$ 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 の特徴は，低次元表現と再構成の誤差をつなげた $\boldsymbol{z} = [\boldsymbol{z}_c, \boldsymbol{z}_r]$ を最終的な潜在表現として利用することである．再構成の誤差が，潜在表現空間での距離に直接影響する．潜在表現のクラスタリングは，典型的な GMM の表記にならい，式 (1) で定義される．ここで，K はクラスタ数，N はデータの数，MLN は Multi Layer Network の略である．また，クラスタ k の混合係数は式 (2)，平均と分散共分散行列は式 (3) となる．

$$\mathbf{p} = MLN(\mathbf{z}; \theta_m), \hat{\gamma} = \mathrm{softmax}(\mathbf{p}) \tag{1}$$

$$\hat{\phi_k} = \sum_{i=1}^{N} \frac{\hat{\gamma}_{ik}}{N}, \tag{2}$$

$$\hat{\mu_k} = \frac{\sum_{i=1}^{N} \gamma_{ik} \boldsymbol{z}_i}{\sum_{i=1}^{N} \gamma_{ik}}, \hat{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^{N} \hat{\gamma_{ik}} \left(\boldsymbol{z}_i - \hat{\mu}_k\right) \left(\boldsymbol{z}_i - \hat{\mu}_k\right)^\top}{\sum_{i=1}^{N} \hat{\gamma}_{ik}} \tag{3}$$

ある入力 $\boldsymbol{x}$ の潜在表現 $\boldsymbol{z}$ について，これが異常である度合いは，式 (4) のエネルギー関数の値であらわされる．これは，直感的には，k 番目のクラスタの中心 $\hat{\mu}_k$ から $\boldsymbol{\Sigma}_k$ を用いて $\boldsymbol{z}$ への距離を測り，全クラスタからの距離の和が大きい $\boldsymbol{z}$ を異常と判定していると解釈できる．もちろん，「異常」の比率はデータに依存する．[10] の例では，単純に，エネルギー値上位 20% を異常と判定している．

$$E(\boldsymbol{z}) = -\log \left(\sum_{k=1}^{K} \hat{\phi}_k \frac{\exp\left(-\frac{1}{2} \left(\boldsymbol{z} - \hat{\mu}_k\right)^\top \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{z} - \hat{\mu}_k\right)\right)}{\sqrt{\left|2\pi \hat{\boldsymbol{\Sigma}}_k\right|}} \right) \tag{4}$$

訓練は，ニューラルネットワークのパラメタ $\theta_e, \theta_d, \theta_m$ に対して，下記の目的関数を最小化することで行う．L はベクトルの再構成に関する損失関数，P は罰則項であり，λ はハイパーパラメタである．

$$J(\theta_e, \theta_d, \theta_m) = \frac{1}{N} \sum_{i=1}^{N} L(\boldsymbol{x}_i, \boldsymbol{x}_i') + \frac{\lambda_1}{N} \sum_{i=1}^{N} E(\boldsymbol{z}_i) + \lambda_2 P(\hat{\Sigma}) \tag{5}$$

4. 実験結果

イギリス英語母語話者による英語の均衡コーパスとして，代表的な British National Corpus（BNC）のうち，10 万文に対して BERT[2] 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

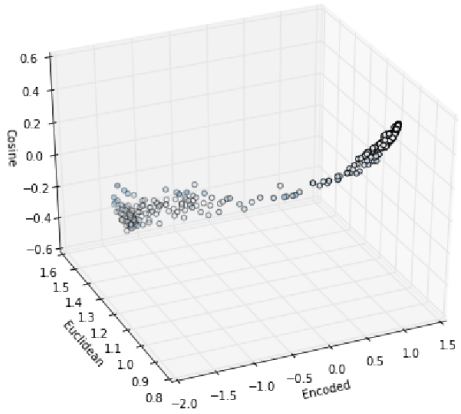

図 1 “period” の 3 次元の可視化．色が濃いものが外れ値と判定されている．この図を真上から見た図が図 2，図 3 である．

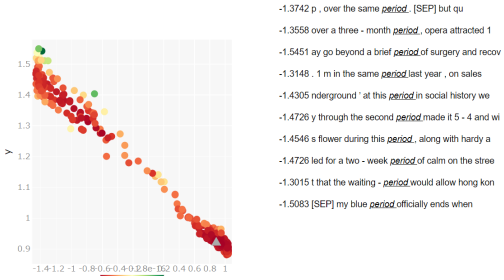

図 2 “period” の主要な用例．丸い各点は，“period” の各用例（コーパス中の各出現）に対応する．大まかには意味的に近い点同士が図上でも近く表される．各点の色は例外的である度合い（エネルギー値）を表し，緑色ほど例外的，赤色ほど主要と判定されている．強い緑色は外れ値と判定され，十字で表される（閾値は自動調整される）．右下の赤い点が多く集まる部分にある，灰色の▲が基準点であり，基準点からの距離が近い点 10 点に対応する用例が，テキストの形で右側に示されている．テキストの前の数値は，実際の各用例のエネルギー値である．本稿の用例は，全て BNC[1] から取得した．

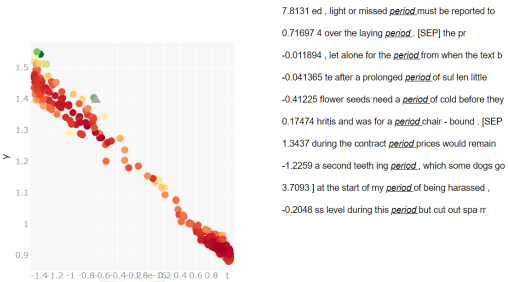

図 3 “period” の例外的な用例．例外的と判定された緑色の点に合わせて基準点を設定し，この緑色の点に対応するテキストが，右側の一番上に表示されている．

[*1] https://github.com/huggingface/transformers

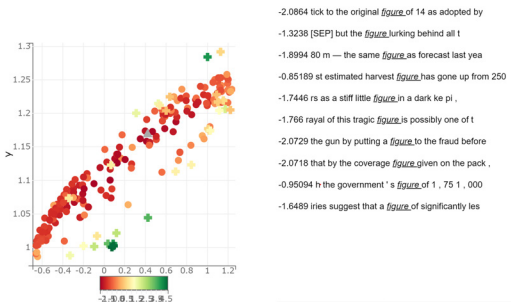

図 4 “figure” の主要な用例その 1．丸い各点は，各用例（コーパス中の各出現）に対応する．▲を濃い赤の部分（主要）な用例に合わせた．テキスト中の数値からも，主要な用例と判定されていることが分かる．数値（桁数）の意味の “six-figure mile”，人物の意味の “child-like figure”，「人目を引く」という慣用句の “cut a figure”，など，主要と判定された，多義語 “figure” の用例が意味的に多様であることが分かる．

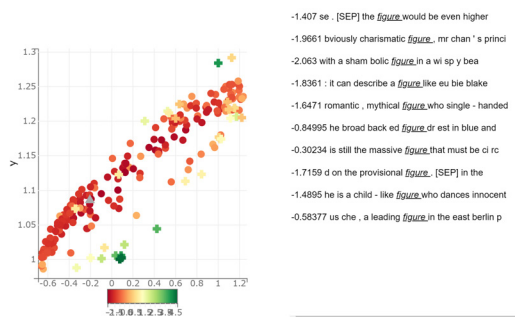

図 5 “figure” の主要な用例その 2．▲を濃い赤の部分（主要）な用例に合わせた．テキスト中の数値からも，主要な用例と判定されていることが分かる．

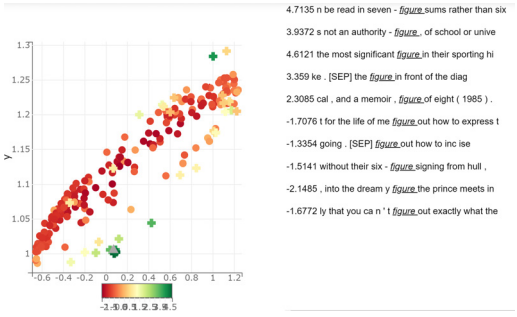

図 6 “figure” の例外的な用例．▲を濃い緑の部分（例外的）な用例に合わせた．テキスト中の数値からも，例外的な用例と判定されていることが分かる．

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，次元数の他は，この実装で用いられているものと同一とした．特に，DAGMM のクラスタ数は 4 と設定されている．

[*2] https://github.com/danieltan07/dagmm

[13] との比較のため，“period” という語を例に議論する．**図 1** に，DAGMM による用例の潜在表現 z を 3 次元で表示した．ここでは，色が濃いほど，エネルギー値が高く，外れ値であると判定されている．紙面での見やすさを考慮して．z のうち最初の 2 次元分を用い，図 2 と図 3 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．各点の色は式 (4) であらわされる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例外的ではない，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解釈できる．

横軸・縦軸は，それぞれ，DAGMM の潜在空間表現 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，10 点を並べ，これに対応するテキスト 10 件が，用例として右側に提示されている．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 2 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，「はじめに」で説明した，1 の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．可視化・クラスタリングの観点では有用でない結果であっても，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，2 の機能の観点では有用であろう．図 2 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 2 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，異常度が低い語を，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．

4.1 “period” の例

図 3 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．

生理の意味での “period” が例外的と判定されたのは，単純に，このコーパスの中でその意味で使われることが少なかったからであると思われることに注意されたい．提案手法は，対象とするコーパスの意味的な外れ値を探しているだけである．逆に，医療コーパスのように，医学的な意味での単語が主要なコーパスを対象にすれば，生理の意味の “period” が主要な結果になることもあり得る．このように，医療を英語で学ぶ学生のための語彙学習支援，というような，非常に specific な要件であっても，その要件を満たすコーパスが学習対象言語中に存在していれば，適応的に支援することが可能であることが提案手法の利点の 1 つである．

最後に，各単語の異常度の閾値をパラメタとして，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 [14] を実装し，間接的な精度評価を行った．100 語種について 100 人をテストした単語テストデータ [3] 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，異常度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 2 や図 3 に示す詳細な分析が可能であることが示された．

4.2 “figure” の例

“period” は，時代という日本語では別の言葉で表される場合があるにせよ，ほとんどの場合，ある長さの時間（期間）を表しており，BNC という均衡コーパス中では，「生理」という意味はごく少数であった．もう一つの例として，“figure” の例を示す．“figure” も英語学習者が初期に学ぶ重要単語でありながら，主要な意味に「数字」や「人物」などの様々な意味が現れる単語である．どの意味も同程度に重要なため，語彙学習支援の場においては，語彙曖昧性解消などと異なり，主要な用例として，どの意味も現れてほしいという状況がある．

図 5 に，参照点▲を濃い赤の部分に異動させ，“figure” の主要な用例のリストを表示させた．この図では，以前の図と少し異なり，例外的と判定された用例（$E(z)$ が閾値以上である用例）を “+” 表記で表すことで，例外であることを明示している．用例の前に示される数値が正の場合が例外的，負の場合が主要と判定されている．この例では，“mythical figure”（神話上の人物）や，“charismatic figure”（カリスマ的な人物）のように，人物としての “figure” が▲の近くに表示され，主要な用例として提示されている．一方，“provisional figure”（暫定的な数字）というように，数値の意味での “figure” も主要な用例として提示されており，人物の意味での “figure” と数値の意味での “figure” どちらも主要な用例として判定されていることが分かる．図 5 では，どちらかというと人物の用例が多いように見受けられ，人物の用例が固まっていることが示唆される．

図 4 では，参照点▲を別の濃い赤の部分に異動させることで，他の主要な用例を提示させている．ここでは，“the same figure as forecast”（予想と同じ数値）や，“estimated harvest figure”（推定される収穫量）のように，数値の意味での “figure” が固まっていることが分かる．一方，“figure lurking behind”（〜の後ろに潜む人物）のように，人物の意味の “figure” もあり，どちらの用例も学習上重要である判定されていることが分かる．

最後に**図 6** に，例外的と判定された用例の “figure” を示す．この例では，“seven-figure” や “authority-figure” のように，単純に珍しい使われ方をしている用例が例外的と判定されている．これは，BERT も近隣の語の出現パターンから文脈化単語埋め込みベクトルを生成するので，単純に単語の組み合わせがなかなか出てこない用例が例外的と判定されているものと思われる．“seven-figure” が例外的であるのに対して，その下の “six-figure” は主要と判定されているが，これは “six-figure salary” や “six-figure income” という，年間の給与額が 100,000 米ドル/ポンド以上の高給を表す表現が米国/英国に存在するため，“six-figure” という単語のパターンの頻度が “seven-figure” より高く，主要と判定された可能性がある．

5. 分析とデータセット作成

図 2 は，2 つのクラスタに分かれている．この可視化が，各用例の語義を反映している事が保証できれば，学習者にとって有用であろう（「はじめに」の機能 1）．しかし，人手の語義のアノテーションを用いない手法である以上，各クラスタが語義によって分かれていることを保証することは難しい．クラスタリング結果が必ずしも語義を反映していなくとも，学習の優先度が高い用例であれば学習すべきことに変わりはなく，主要度の表示は機能 2 の観点では学習者に有用であろう．図 2 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 2 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．

図 3 に，基準点を例外的な用例に合わせた例を示す．右側には “light or missed period” という用例が表示されている．“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，BNC の中では使用例が少ないためか，例外的な用例と判定され

表 1 各語の例外的/主要な用例のリスト

語	単語頻度	主要な用例	例外と判定された用例
time	2,863	2,753	110
see	1,359	1,288	71
period	376	368	8
poor	275	269	6
deficit	137	136	1
restore	53	52	1
olive	43	41	2
ubiquitous	13	13	0
retro	13	13	0
weep	8	8	0

ていることがわかる．また，この用例の “period” は名詞であり，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など他の自然言語処理技術を用いて捉えることは難しい．“period” のこの 2 用例の比較のため，複数の英語母語話者の確認の元，両者の意味を問うテストを作成し，クラウドソーシング上で 235 人の被験者から回答を得た．期間の “period” の正答者は 180 人，生理の “period” の正答者は 76 人であり，後者が明らかに難しいことから，異常度が低い用例が直感的にも語の主要な用例であることが質的・量的に示された．全体で 13 語の主要/非主要な語義を各々学習者に問い，平均正答率は主要語義で 68.2%，非主要語義で 39.8%であった．このデータや詳細な関連実験については，`http://yoehara.com/` に公開予定である．

表 1 に，提案手法により判定された，各語の用例のうち主要な用例の単語頻度を記す．提案手法における各単語の異常度の閾値は，学習者に対する語彙テスト結果データ（100 語種について 100 人をテストしたもの [3]）を用いて自動的に調整した．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，学習者の語彙テストの正答/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，異常度を用いた頻度修正を行った特徴量を用いてロジスティック回帰で予測したところ，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 2 や図 3 に示す詳細な分析が可能であることが示された．

6. おわりに

本稿では，語の用例を，多義性を考慮して視覚的に提示できる，教師なし深層異常検知に基づく外国語の語彙学習支援 UI を提案した．提案手法は，修正前後の用例数を特徴量に用いた精度評価では従来手法と同等精度でありながら，主要な用例，例外的な用例を適切に認識できる事を，データセット作成を交えた評価で示した．

提案手法は教師データを必要としないため，様々なドメインや言語に適用できる利点がある．例えば，科学論文の中で使われる “figure” という語の外れ値なども求めることができる．今後の課題としては，他ドメインや多言語への展開が挙げられる．

謝辞 本研究は，科学技術振興機構 ACT-X 研究費 (JPMJAX2006)，ならびに日本学術振興会科学技術研究費補助金（18K18118）の支援を受けた．また，産業技術総合研究所の AI 橋渡しクラウド (ABCI) を使用した．

参考文献

[1] BNC Consortium: *The British National Corpus* (2007).
[2] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. of NAACL* (2019).
[3] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).
[4] Ehara, Y.: An Approach to Summarize Concordancers' Lists Visually to Support Language Learners in Understanding Word Usages, *Proc. of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI)*, (online), DOI: 10.18653/v1/W19-8407 (2019).
[5] Ehara, Y.: Semantically Adjusting Word Frequency for Estimating Word Difficulty from Unbalanced Corpora, *Companion Proc. of LAK* (2020).
[6] Luo, Y., Zhao, H. and Zhan, J.: Named Entity Recognition Only from Word Embeddings (2019).
[7] Paetzold, G. and Specia, L.: SemEval 2016 Task 11: Complex Word Identification, *Proc. of SemEval-2016*, (online), DOI: 10.18653/v1/S16-1085 (2016).
[8] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep Contextualized Word Representations, *Proc. of NAACL-HLT*, (online), DOI: 10.18653/v1/N18-1202 (2018).
[9] Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A. and Zampieri, M.: A Report on the Complex Word Identification Shared Task 2018, *Proc. of BEA* (2018).
[10] Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D. and Chen, H.: Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection, *Proc. of ICLR*, (online), available from ⟨https://openreview.net/forum?id=BJJLHbb0-⟩ (2018).
[11] 中田達也：英単語学習の科学，研究社 (2019).
[12] 江原遥：文脈を考慮した視覚的な語彙学習支援，NLP 若手の会第 14 回シンポジウム (2019).
[13] 江原遥：文脈化単語表現空間上の範囲の学習による語の多義性を考慮した頻度計数法，第 243 回自然言語処理研究発表会予稿集 (2019).
[14] 江原遥：深層異常検知に基づく多義語のコアミーニングを考慮した既習語予測モデルの定式化，言語処理学会年次大会予稿論文 (2020).
[15] 江原遥：語彙学習支援システム（私のブックマーク），人工知能学会誌，Vol. 35, No. 2, pp. 296–300 (2020).
[16] 芦原和樹，梶原智之，荒瀬由紀，内田諭：多義語分散表現の文脈化，自然言語処理，Vol. 26, No. 4 (2019).
[17] 鈴木類，古宮嘉那子，浅原正幸，佐々木稔，新納浩幸：概念辞書の類義語と分散表現を利用した教師なし all-words WSD，自然言語処理， Vol. 26, No. 2 (2019).