優先的に学習すべき語の用例の視覚的な自動推薦

江原 遥
Yo EHARA
静岡理工科大学情報学部
Department of Informatics, Shizuoka Institute of Science and Technology
Email: ehara.yo@sist.ac.jp

あらまし： 外国語学習の語彙学習においては、従来、単語テストの結果から、学習者が学習すべき語を自動的に提示して推薦する人工知能技術はあった。しかし、語には複数の意味を持つ多義語が多くあり、各意味ごとに代表的な用例がある。本稿では、母語話者生コーパスと学習者の単語テスト結果のみから、入力された語について、学習者が知らないと推測される代表的な用例を自動的に判定し、学習者に推薦する手法を提案する。

1 はじめに

本稿では [10] を要約する．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 [6, 3]，語義曖昧性解消にもすでに利用されている [8]．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」[7] 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 [3] と教師なし深層異常検知 [7] 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．評価については，語の各語義について，学習上の優先度を人手で付与した高品質なデータセットが知る限り存在しないので，例を通じた質的評価と間接的な量的評価を行った．

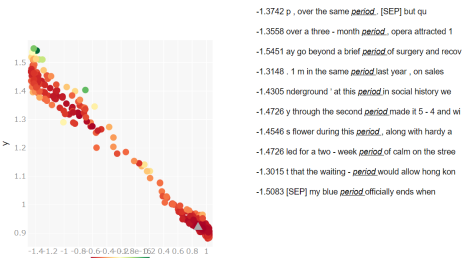

図 1: “period” の主要な用例．丸い各点は，“period” の各用例（コーパス中の各出現）に対応する．各点の色は例外的である度合い（エネルギー値）を表し，緑色ほど例外的，赤色ほど主要と判定されている．右下の赤い点が多く集まる部分にある，灰色の▲が基準点であり，基準点からの距離が近い点 10 点に対応する用例が，テキストの形で右側に示されている．テキストの前の数値は，実際の各用例のエネルギー値である．本稿の文例は，全て BNC[1] から取得した．

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[7] 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 [8]．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 [5]．

DAGMM は，入力ベクトル $\vec{x}$ をオートエンコーダ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 の特徴は，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．

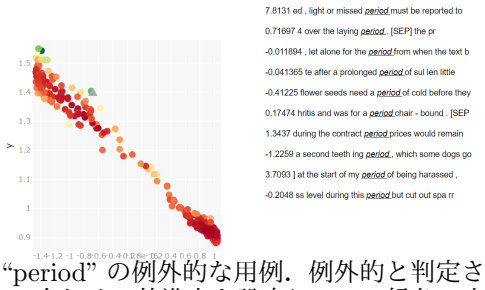

図 2: “period” の例外的な用例．例外的と判定された緑色の点に合わせて基準点を設定し，この緑色の点に対応するテキストが，右側の一番上に表示されている．

3　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[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．

[9] との比較のため，“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 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 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，「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 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，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．

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 [11] 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 [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，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 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．

4　おわりに

本稿では，語彙学習者の目的で，語の用例を，多義性や主要性を考慮して提示する，教師なし深層異常検知に基づく手法を提案した．提案手法によって判別された主要な用例は，主要用例数のみを特徴量に用いた間接評価では通常の単語頻度と同等精度でありながら，質的評価で，主要な用例，例外的な用例を適切に認識できる事が示された．今後の課題としては，言語資源を作成し，例外的な用例の詳細な数値的な精度評価を行うことが挙げられる．さらに，[2] などを拡張し，語彙学習を強化学習の観点からモデル化することも面白い．

謝辞

本研究は，科学技術振興機構 ACT-I 研究費 (JPMJPR18U8)，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．

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

参考文献

(1) BNC Consortium. *The British National Corpus*. 2007.
(2) Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie. DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, May 2019. arXiv: 1905.06873.
(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) Ying Luo, Hai Zhao, and Junlang Zhan. Named entity recognition only from word embeddings, 2019.
(6) Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
(7) 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.
(8) 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.
(9) 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.
(10) 江原遥. 外国語語彙学習のための教師なし深層異常検知に基づく語の用例の多義性・主要性の提示. 人工知能学会年次大会予稿論文, 2020.
(11) 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.