公開

（事後評価・研究者用）

研　究　報　告　書

「意味空間上の広がりに基づく効率的な語彙学習支援システム」

研究期間：2018 年 10 月～2020 年 3 月
研究者番号：50168
研 究 者：江原　遥

1．研究のねらい

外国語の語学学習は、他の学習と異なり、外国語を用いて、語学以外の他の専門分野の習得するニーズが高いことが特徴である。例えば、英語を用いて科学を勉強する場合がこれに当てはまる。このように、他の専門分野を習得するために語学を学ぶ場合には、各専門分野のテキストが読解できるかどうかが、最終的なゴールとなる。

このように、他の専門を学ぶために必要な語学学習を考えると、学習者にとっては、まずは、専門分野のテキストが読みこなせること、すなわち「読解」が重要となる。作文・会話・聴解は重要であるが、読解できない内容を作文・会話・聴解する事は難しいと考えられるので、まずは、内容を読解できる段階に到達することが必要であろう。

専門分野によらず、読解のためには語彙学習が重要であることが既存研究で知られている。文法などの基礎的な事項を理解していても、テキスト中に現れる語のうちの一定割合以上を知らなければ、テキストの内容を十分に理解することができないことが確かめられている。

このように専門分野の習得まで含めた理想的な外国語の語彙学習を考えると、各学習者に適応して、学習者ごとに覚えるべき語や語の用例が提示できる学習支援が望ましい。学習者ごとに、習得すべき語彙も、現在の習得状況の評価も異なるからである。このような学習支援は、一般に、人間の語学教員だけで実現することは困難である。一般に人間の語学教員は、学習者の専門分野については知識がないこと、語彙学習は長時間にわたるため語学教員を拘束し続けることが高コストだからである。学習者の自宅など、場所を選ばない遠隔環境で、学習者の目的・現状に合わせて、語や語の用例が学習事項として推薦される仕組みの方が現実的かつ低コストであろう。

本研究では、専門分野の要学習語やその文脈ごとの使い分けを、学習者に合わせて提示・可視化する、効率的な語彙学習支援システムを開発することを狙いとする。技術的には、語が持つ文脈の多様性を空間的な広がりとして表現可能な、分散表現の獲得手法を開発する。

語学学習を通じて専門分野の習得を行う場合には、単に学習中の言語の語彙と母語での意味を暗記するだけでは不十分で、専門分野ではどのような文脈（用例）があり、各々の語をどのような文脈でどのように使い分けるのか、という細かい使い分けの知識が必要となる。例えば、工学分野の習得を目的とする英語学習において、”tremble”, “vibrate”, “oscillate”という語の学習を考えた場合、単にこれらが「震える」という日本語の意味に対応することだけではなく、”tremble”は主に体の震えなどの表現に使い工学の振動を表現する目的では後者の2語が使われる、といった知識を習得する必要がある。こうした使い分けの知識を専門分野と英語の両方を解する専門家が書き下して教材を作成することは、経済的・時間的に非効率的である。

このような、専門分野ごとの使い分けの知識を自動的に抽出するため、本研究では語の「広がり」を持った分散表現の空間（意味空間）を考える。具体的には、個々の使い分けを表現する学習事項（用例）を特定し、従来のように語ではなく、用例を点とする分散表現空間の構築を目指す。語はこの空間上の「広がり」を持った領域で表現される。語と語の使い分けの差は、領域の重なり具合で判断される。学習者の専門分野上で重要な用例や、学習者の語彙能力もこの空間上で表現する。これにより、学習者の能力に合わせて学習者が志望する専門分野で重要な語やその用例を学習者に提示し、空間を可視化することで効率的な個人化語彙学習支援を実現する。

2. 研究成果

（1）概要

本研究の成果は、「語彙学習のための領域表現」と、「語彙学習者が読解可能なテキストの個人化判別」の2つの研究成果に大別される。どちらでも、新規な技術を開発し、査読付き国際会議や国内会議の受賞により、対外的に高い評価を得た。次に、1つずつ説明する。

「語彙学習のための領域表現」では、語彙学習者が用例の使い分けを理解可能にするための領域表現と可視化の研究を行った。具体的には、次の2種の新規技術の研究に従事した。

研究テーマ 1-1） 語彙学習者が語の使い分けを学ぶためには、まず、各単語に、自分が知っている用例以外にどのような用例があるのかを、直感的に確認できることが重要である。この目的のため、ある単語の各用例を点として、用例間の意味的近さが図上の距離に対応するような図を出力する技術を研究した。この図はインタラクティブ性をそなえており、学習者は、図上の各点が実際にどのような用例かを閲覧できる。

研究テーマ 1-2） 研究テーマ 1-1 により出力される図で、各学習者が知っていそうな用例の範囲があらかじめ推定されていれば、学習者は、より素早く自分が未学習の重要な用例を探し出すことができる。そこで、各学習者が学習済みの用例を推定する技術を開発した。この技術により、意味的距離が多少離れていても、多くの学習者が知っている用例はかたまって表示されるようになり、学習者が未学習の用例をより素早く見つけられる。

次に、「語彙学習者が読解可能なテキストの個人化判別」の研究の動機について説明する。本研究は、語学を通じて何か他の専門を学ぶような語学学習者を想定しているが、学習者はどのように専門分野を指定すれば便利だろうか？専門分野が細分化されている今、様々な専門を網羅する分類体系を構築することは困難であり、本研究の範囲を超える。そこで、本研究では、専門分野の分類体系を構築するのではなくて、学習者に、読みたい本やテキストを具体的に提示してもらう事で、専門分野の指定を行うことにした。この目的のため、次の技術を開発した。

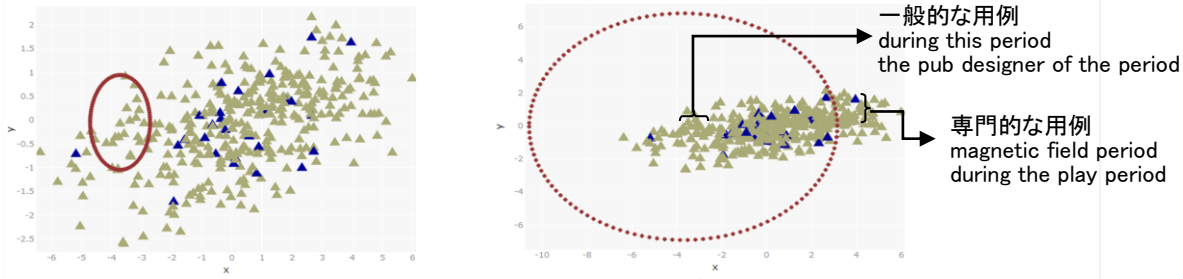
研究テーマ 2-1）語彙テストの結果だけから、各学習者が、所与のテキストを読解可能か判定する技術の開発

（2）詳細

研究テーマ 1-1 では、語彙学習者に、どのような用例があるのかを可視化を通じて直感的に確認できるようにするために、ある単語の各用例を点として、用例間の意味的近さが図上の距離に対応するような図を出力する技術の研究を行った。図 1 に、具体的に得られた図を載せる。この図はインタラクティブ性を備えており、ユーザは用例間の意味的近さを確認しながら、図の各点

に対応する用例を表示させることができる。

こうした図を出力するためには、用例間の意味的近さを計測する必要がある。この意味的近さの数値を膨大な種類の語に対して人手で付与することは、コストが高く非現実的であるうえ、新語や専門用語には対応しづらい技術的課題がある。この課題に対し、本研究では、母語話者の書いたテキスト集合（コーパス）中の、語の1出現を1用例とみなし、文脈を考慮して語の出現を可視化することで解決した。文脈を考慮した語の出現の意味の可視化については、語の出現ごとに、文脈を考慮した語の埋め込みベクトル表現を求める「文脈化単語埋め込み」[Devlin et al.,2019]の技術を用いた。



“period”の用例の推定（学習）された可視化．▲は用例．青い（濃い）▲は科学分野のテキスト中に出てきている用例．円内の用例が，各学習者に知られていると判定された．

図 1:用例の可視化、学習者が知る用例範囲の推定

研究テーマ1-2では、研究テーマ1-1に加えて、学習者が知っている用例の範囲を自動的に推定する機能を持たせることにより、学習者にとって学習優先度の高い用例をより容易に見つけられる可視化技術を提案した（図 1）。従来、学習者が知っている語彙は、数十分で回答可能な100単語程度の単語テストを受けてもらえれば、高精度に推定可能であることが示されている。一方、用例は1単語に対して多数存在するため、学習者に用例のテストを受けてもらい直接的に計測する方式では、回答に数十倍～数百倍の時間がかかることが予想され、学習者に過大な負担がかかる問題がある。この問題に対処するため、本研究では、従来使われてきた数十分で回答戒能な単語テストデータのみを用いて、図中で学習者が知っている用例の範囲を推定可能なモデルを提案した。

この提案における研究上の主たる課題は、各用例を点として各用例の意味的近さを表す図中の用例点集合と、単語テストにおける単語の難易度を、ヒューリスティクスではなく理論的に正当性のある形で結びつける事である。この課題に対して、用例点集合の総数はコーパス中の単語の出現数、すなわち単語頻度であることに着目した。そして、範囲中の用例点数を単語の難易度とみなし、単語テストデータにフィットするような用例点数を返す図上の範囲を推定するニューラルモデルを構築する問題に帰着することで、この課題を解決した。副次的な技術的課題として、範囲中の「用例点数」は離散値であり、ニューラルモデルの訓練に通常用いられるライブラリではモデルパラメタの最適化が行いづらい。この技術的課題に対処するため、提案手法では、新規な連続緩和も提案している。機械学習の用語を用いて、提案したニューラルモデルを説明すると、大規模な母語話者コーパスからの事前学習により構築された文脈化単語埋め込みによって得られる各単語の用例点集合の散らばり具合の情報を、単語テスト結果データにおける各学習者の反応予測問題を解く際の単語難易度の情報として、転移学習していることに相当する。

研究テーマ 1-1 は、NLP 若手の会第 14 回シンポジウム奨励賞を獲得し、NL4XAI ワークショップに採択された（主な研究成果 3）。研究テーマ 1-2 は、ラーニングアナリティクスのトップ国際会議である LAK に poster paper として採択された（主な研究成果 4）。この他、研究テーマ 1-2 につながる、単語埋め込みと単語難易度の関係性に関する論文が PACLING の poster paper に採録されている（主な研究成果 5）。

研究テーマ2では、語彙テストの結果だけから、各学習者にとって所与のテキストが適度な難度を持っているか判定する手法を提案した。具体的には、各学習者にとって所与のテキストがどの程度読めないのかを確率として求める手法を提案した。既存には、テキスト中で学習者が知っている語の比率である既知語率（テキストカバー率）が 95%以上でなければ、学習者はテキストを読めないという研究結果が多数の研究により報告されていた。しかし、この既存手法では、判定結果は読める/読めないの 2 通りであり、語彙学習に重要な「惜しいケース」がどの程度惜しいのかが評価できない。例えば既知語率が 90%と推定された学習者は、実際には全くテキストが読めないわけではなく、何回かに一回はテキストを読みこなせる程度には能力があると思われるが、この「何回かに一回」の具体的な数値は評価できなかった。

この技術的課題に対処するため、既知語率自体を確率変数としてとらえ、「既知語率が 95%以上になる事象」の確率値を、単語テストデータから計算する手法を提案した。技術的には、この問題がポアソン二項分布の累積確率分布を求める問題に帰着することを示し、そのための動的計画法に基づくアルゴリズムを提案した。これにより、学習者がテキストを読める程度が確率値として評価できるようになった。実際に、単語テストと読解テストを同一の被験者に回答してもらったデータを作成し、このデータを用いて提案手法の性能を評価した。提案手法は、語彙学習者が単語を知っているかどうかの確率値の情報を利用できる分、読解テスト結果の予測性能は提案手法が統計的有意に上回った。さらに、提案手法は、図 2 のように、ある学習者があるテキストを読んだ際の既知語率の確率分布を出力することを可能にしており、単にテキストが読める/読めない、ではなく、テキストが読めない場合の「惜しさ」を確率という尺度を用いてテキスト間/学習者間で比較可能な形で提示することができた。

研究テーマ2は、第 44 回教育システム情報学会全国大会で大会奨励賞を受賞した（主な研究成果 1）。また、機械学習応用の査読付き国際会議である ICMLA2019 に採択された（主な研究成果 2）。

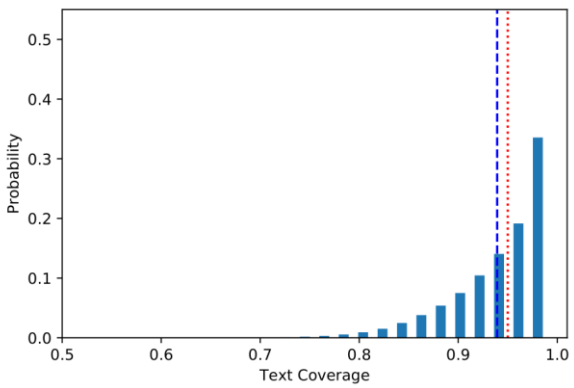


図 2：研究テーマ 2。ある学習者のあるテキストに対する既知語率の確率分布。

3. 今後の展開

研究テーマ 1-1、および、研究テーマ 2-2 については、語彙学習者の学習に適した用例可視化という、新規な研究シーズを掘り起こしており、学術上の高い波及効果が期待できる。当初の狙いは、複数の語の用例を 1 つの図上で表示することにより、似たような語の意味の「使いわけ」を学習する事であった。これを適切に行うためには、まず、1 つの語の各用例を、意味を捉えながら適切に可視化することが基本となる。そのため、本研究では、まずは、1 語の用例の可視化の研究に注力した。複数語の用例可視化は、今後の課題である。また、本研究は、そもそも、語彙学習者の自習による語学学習を念頭に置いているため、今後、社会的に強いニーズが生じると思われる遠隔講義による語学学習支援の方面でも、幅広い教育システム上の応用があるものと期待される。

1 語の用例の可視化についても、評価や手法の面では課題がある。学習者が知っている「用例」の特定については、現状、訓練用のデータも評価用のデータも存在しない。訓練については、今回 1-2 で提案しているモデルのように、事前学習を用いた文脈化単語埋め込みモデルからの転移学習を用いて、用例の意味的な差異の程度情報を転移させることが基本的技術になる。また、用例点集合と単語難易度を結びつける「範囲」の考え方にも進展がある。例えば、多義語については、用例点集合が複数の塊（クラスタ）に分かれて存在しているため、今回の研究のように、範囲は円形とする仮定になじまない技術的課題がある。直近では、この技術的課題に対し、単純に円の範囲でカウント対象の用例点集合を分ける、教師なし深層異常検知モデルを組み込み、各用例点に対して算出される「異常度」の閾値を用いてカウント対象の用例点集合を分ける手法を提案している。具体的には、教師なし深層異常検知モデル DAGMM [Zong+,ICLR2018]を用いて、文脈化単語埋め込みベクトル可視化次元への潜在表現への射影，射影空間での混合ガウスモデル（Gaussian Mixture Model）によるクラスタリング，どのクラスタからも遠い用例点は例外的な用例として学習優先度を低く設定する。実際に、国内会議で 2 件こうした研究について発表している。

研究テーマ2の今後の進展としては、次の課題が挙げられる。研究テーマ2では、各学習者が各テキストを読める度合いを確率値として表現することにより、異なるテキスト間・異なる学習者間で比較可能にした（図 1）。しかし、この確率値を用いた表現は新しいので、語彙学習に適した難度が、具体的に、提案手法において何%の確率であるのかは、語彙学習者を用いたユーザ実験を通して検証する必要がある。

また、語彙学習の観点からおそらくニーズが高い機能として、具体的にテキスト中のどの部分の単語を知っていれば、そのテキストが読めるようになるのかを出力する機能が挙げられる。この機能も、今回提案した動的計画法によるアルゴリズムの延長で実現可能ある。しかし、実際にこの出力が正しいのかを評価するためには、やはり、評価用データの作成が重要になろう。

4. 自己評価

研究目的の達成状況： 新しい研究分野を開拓するという点においては、期待以上の成果があったと評価する。語彙学習支援のための新規なタスクを提案し、そのタスクのプロトタイプとなる機械学習モデルをも設計できた。当初の目的であった、「語の意味的な広がり」を、語の用例集合の点集合で表現する技術、学習者が知っている語の用例の範囲を推定する技術も提案した。これに加えて、語彙学習者がテキストを読めるか個人化判定する技術も開発した。

一方で、新規なタスクに対するモデルを提案してきたことから、提案に対して評価・検証用のためのデータ作成が追いついていないことは確かである。例えば、学習者が知っている語の用

例を実際に記録したデータや、作成した可視化が学習者にとって本当に有用かのユーザ実験データなどが新たに必要となろう。

研究の進め方： 研究期間中に、「文脈化単語埋め込み」という新規技術が誕生し、自然言語処理分野の主流手法全体がこの技術を用いたものに変わっていくという、非常に大きな技術的ブレイクスルーがあった。このブレイクスルーをうまく活用し、多くの査読付き国際会議論文を出版できたため、総合的にはおおむね順調であったと評価する。

研究の波及効果： 語彙学習支援のための新しいタスクを掘り起こした点で、学術上の波及効果は高いと思われる。これで検証用のデータを作成し、ちゃんとタスクの重要性を示すことができれば、どれもトップ査読付き国際会議の Full paper 数本分の内容であると考えている。今後、継続的に関連研究を投稿することを目標としている。2020 年度の言語処理学会や人工知能学会における発表は、そのための準備である。また、アカデミアへの波及効果として、本研究の人工知能学会学会誌の「私のブックマーク」において、語彙学習支援システムに関する特集解説記事を担当させていただいた。

産業への波及効果としては、語彙学習支援システムを実際に作成している企業には、社会実装のための関心を持っていただいている。例えば、英語学習ソフトウェアを作成しているある企業からは、実際に講演のためのコンタクトがあった。このように産業界からも関心を持っていただいているため、今回の研究で新たに作成したデータセットやソースコードを公開していくことで、産業への波及効果を高めていきたい。

また、この分野の研究やソフトウェアは、語学教員が個人で作成しているものも多い。今回開発したインタラクティブに用例を把握できる語彙学習支援のデモシステム（研究テーマ 1-1）や、語彙学習に適したテキストを検索するデモシステム（研究テーマ 2）を公開すれば、語学教員や語彙学習者にも波及効果は高いと思われる。特に、本研究で提案した手法は、どれも言語に大きく依存しない手法であるので、英語以外の言語の語彙学習支援にも容易に対応できる。

研究課題の新規性・独創性は、実際の語学学習者のニーズにあった新しいタスク設定を設定し、実際に、自然言語処理分野と教育分野の双方の国際会議論文として採択されていること、教育分野の国内会議で論文・発表審査の上で受賞していることから、明確に高いと評価する。

技術的にも、研究テーマ2）では、既存手法を数理的に発展させ、過去 30 年間知られていなかった確率的拡張を行っている。また、研究テーマ 1-1、1-2 でも、同年に採択されたばかりの手法（BERT）[Devlin+,2019]を語彙学習支援に応用し、新規のモデル提案に至っている。特に学習者が知っている用例の範囲を深層学習の枠組みで求めるための連続緩和も、独自のものである。

5.　主な研究成果リスト

（1）論文（原著論文）発表

1. 江原遥. テキストカバー率の確率的拡張に基づく語彙テストのみからの個人化読解判定. 第 44 回教育システム情報学会全国大会講演論文集. （論文・発表審査により大会奨励賞受賞）

2. Yo Ehara. Uncertainty-aware Personalized Readability Assessments for Second Language Learners. In Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA).

3. Yo Ehara. An Approach to Summarize Concordancers' Lists Visually to Support Language Learners in Understanding Word Usages. In Proc. of a of the 12th International Conference on Natural Language Generation (INLG 2019) workshops, NL4XAI demo paper.

4. Yo Ehara. Semantically Adjusting Word Frequency for Estimating Word Difficulty from Unbalanced Corpora. In Companion Proc. of the Learning Analytics & Knowledge Conference (LAK) 2020, research poster paper.

5. Yo Ehara. Neural Rasch Model: How Word Embeddings Affect to Word Difficulty? In Proc. of the 16th International Conference of the Pacific Association for Computational Linguistics (PACLING 2019).

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

2019/5/29　国内　口頭　江原　遥

テキストカバー率の確率的拡張に基づく語彙テストのみからの個人化読解判定

第 44 回教育システム情報学会全国大会，浜松

2019/9/13　大会奨励賞受賞．論文・発表審査で 4/74 件

https://www.jsise.org/taikai/2019/

2019/7/31　国内　ポスター　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（YANS）第 14 回シンポジウム（2019），札幌

2019/8/28　奨励賞受賞．投票により 7/93 件．

2020/1/17　国内　論文　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, Vol. 35, No. 2, pp. 296-300 2020/3/1