Vol.2021-NL-251 No.11
Vol.2021-SLP-139 No.11
2021/12/2

第二言語学習者にとっての語の難しさ推定に基づく教師なし自動リーダビリティ判定

江原　遥[1,a)]

概要：自動リーダビリティ判定は、テキストの第二言語学習者にとってのリーダビリティを予測するタスクであり、様々な教育応用が見込まれる。このうち、語学教師などによる人手のリーダビリティの教師ラベルを訓練時に全く用いないで判定する「教師なし自動リーダビリティ判定」に着目する。従来は、BERTなどの大規模な事前学習済言語モデルから取得した、外国語テキストのパープレキシティなどの尺度にもとづき判定を行うアプローチが研究されていた [Martinc et al., CL2021]。本研究では、これと異なり、第二言語学習者にとっての語の難しさを推定し、これを活用するアプローチを提案する。具体的には、実際に第二言語学習者から取得した語彙テストの結果データから、項目反応理論に基づくモデルで能力値や語の難しさを推定し、平均的な能力の学習者が所与のテキスト中の各語を知っている確率値を計算して、この確率値をもとに教師なし自動リーダビリティ判定を行う。語の難しさ推定にロジスティック回帰などの比較的軽量なモデルを用いた場合でも、従来の大規模言語モデルを用いた教師なし手法を超える予測性能を確認した。提案手法は、従来手法より低い計算負荷で、より高い予測性能の教師なし自動リーダビリティ判定が行えた。また、British National Corpus などの均衡コーパス中の単語頻度素性をそのまま用いるアプローチでは予測性能は低く、実際に第二言語学習者から取得した「第二言語学習者にとっての語の難しさ」を計算することが予測性能に寄与していることも示唆された。その他、予測性能の適切な評価手法に関する考察や、語の難しさ推定モデルによる結果の説明性、様々なテキストに対して教師なし自動リーダビリティ判定を行った結果などを報告する。

1. はじめに

リーダビリティ（Readability, 可読性）判定は、知的な読解支援システムや言語クラスのクラス分けテストなど、多くの教育応用に直結するため、第二言語習得にとって重要な基礎的タスクである。一方で、リーダビリティ判定は、教育の専門家や言語教師にとってはコストのかかる作業である。この作業を行うためには、テキストを読み、そのテキストが目標とする学習者にとってどれだけ難しいかを推測して、そのリーダビリティを評価しなければならない。教育専門家が必要とする労力のコストを削減するために、言語学習者のためにテキストの読みやすさを自動的に識別するタスクは、過去、多くの既存研究がある。近年、これらの研究は文献 [1] によって「自動リーダビリティ判定（automatic readability assessment, ARA）」という名前で、整理された。文献 [1] によれば、自動リーダビリティ判定は、人手で付与したリーダビリティのラベルを訓練に用いて判定器を構成する「教師あり」設定と、こうした人手のラベルを全く使わない「教師なし」設定に大別される。

このうち、教師なしリーダビリティ判定は早くから登場していた。初期の研究として有名な、Dale-Chall formula (1948) [2]、Flesch Reading Ease formula [3] (1948), Flesch-Kincaid readability formula (1975) [4] などは、コストのかかる人手で付与されたリーダビリティのラベルを訓練に使用せずにリーダビリティのスコアを計算できるので、文献 [1] の分類では教師なしに分類される。これらの計算式は、テキストが与えられたときに、そのテキストに含まれる単語の平均的な長さなどの単純で表面的な特徴に基づいて読みやすさのスコアを算出する。

しかし、これらの初期の計算式の多くは、母語話者の子供にとってのリーダビリティ判定のために設計されたものである。第二言語学習者を対象とした言語教師による可読性ラベルを付与した評価データセットが整備されてきたのは、ずっと後の 2010 年代である [5], [6], [7]。これらの研究では、自動リーダビリティ判定タスクを教師あり文書識別問題として定式化し、この文書識別問題の性能評価用データセットが提案されている。この識別器の構築では、リー

[1] 東京学芸大学
Tokyo Gakugei University, 4-1-1 Nukuikita-machi, Koganei-shi, Tokyo, 184–8501, Japan
[a)] ehara@u-gakugei.ac.jp

ダビリティラベルとの相関が高い特徴を見つけるための特徴工学（feature engineering）に、多大な努力が払われてきた。

これらの研究の後、Bidirectional Encoder Representations from Transformers (BERT)[8] などのニューラル言語モデル (Language Models, LM) が登場し、多くの NLP タスクで優れた結果を達成した。それ以前の研究では、複雑な特徴を見出す特徴工学に依存していたが、ニューラル言語モデルは、大規模なコーパスでの事前学習 (Pre-training) によって、こうした特徴を取り込むことができると考えられる。特に、ニューラル言語モデルのパープレキシティスコアは、与えられたテキストに対する全体的な流暢さと見なすことができる。そこで、[1] では、パープレキシティなどの言語モデルから取得したスコアが第二言語学習者のテキストの読みやすさを表すのにも利用できると推測し、教師なしの自動読みやすさ評価に利用することを提案している。

しかし、このアプローチは、次の問題がある。まず、ニューラル言語モデルにおけるパープレキシティなどの指標は、本質的に母語の「流暢さ（fluency）」の指標であり、これが第二言語学習者にとってのテキストの読みやすさと相関するとは限らない。次に、ニューラル言語モデルを用いることから、本質的に大きな言語モデルを必要とするため、多大な計算量を必要とする事である。

本研究では、リーダビリティ判定のために、第二言語学習者の言語知識を直接測定して識別器を構成するという、これまでとは異なるアプローチを提案する。教育専門家が作成したリーダビリティのラベルは高コストであるので、訓練に用いない「教師なし」設定でのリーダビリティ判定手法を提案する。本研究では、従来軽視されていた第二言語学習者にとってのテキストの難易度の計測データを活用する。具体的には、まず、第二言語学習者に対する語彙テスト結果データから、テキスト中に登場する単語の**平均的な第二言語学習者にとっての**難易度を計算するモデルを構築する。次に、これを用いて、テキスト全体の難易度を計算する。

これまでのアプローチと提案するアプローチの概要を説明する。ニューラル言語モデルに依存する従来のアプローチとは対照的に、提案手法では第二言語学習者にとっての単語難易度の情報を格納している語彙テストの結果に着目し、それを自動リーダビリティ判定に活用する。語彙テストの被験者はテキストの読者である必要はなく、リーダビリティ判定タスクと語彙テストは全く独立である。従来のアプローチも我々のアプローチも、教育専門家が作成した高価な読みやすさラベルを使用していないため、これらのアプローチは「教師なし」自動リーダビリティ判定に分類される。提案するアプローチは、単純な語彙特徴のみを使用するため、ニューラル言語モデルを用いるアプローチよりもはるかに少ないメモリで高速に動作するリーダビリティ判定器を構築できる。実験では、コストのかかるリーダビリティのラベルを訓練に使わず、語彙テスト結果データのみで訓練した提案する判定器は、精度と計算資源の両方において、大規模なニューラル言語モデルを用いたパープレキシティに基づく評価器よりも優れた性能を示した。

訓練	手法	研究の例	可能な性能評価指標
教師あり	ラベル予測	[5], [6], [7], [9], [10]	識別精度、順位相関
	ランキング学習	[11]	順位相関
教師なし	回帰スコア	[12]	
	言語モデルスコア	[1]	
	個人化リーダビリティ	[13], [14], [15], [16], [17]	

表 1 リーダビリティ判定の手法の問題設定の分類。教師あり/教師なしは、各テキストの難度の正解ラベルを用いるか否かを表す。

この論文の貢献は、以下のようにまとめられる。

(1) 第二言語学習者にとってのテキストのリーダビリティ判定のために、軽量な、人手によるリーダビリティのラベルを使用しない「教師なし」の判定器を提案した。
(2) 教師なしリーダビリティ判定器の構築のため、第二言語学習者の語彙テストの結果を利用する方法を提案した。
(3) 公開されている信頼性の高いデータセットを用いた評価実験において、提案する教師なしアプローチは、ニューラル言語モデルを採用した従来の教師なしアプローチよりもリーダビリティの予測性能が高かった。さらに、提案手法は、従来手法よりも省メモリ・高速に動作した。
(4) リーダビリティ判定の評価尺度として、単純な相関係数よりも順位相関の方が優れていることを、スコアの線形性に対する頑健性を通じて示した。

2. 関連研究

自動リーダビリティ判定に関する関連研究を、機械学習の観点から分類し、**表 1** にまとめた。以下、機械学習の観点からのタスク設定の分類ごとに、関連研究を簡潔に紹介する。

ラベル予測は、現在の自然言語処理では典型的な問題設定であり、リーダビリティ判定を、教師あり多値識別問題に帰着させる。代表的な研究としては、文献 [5], [6], [7], [9] など多くの既存研究が挙げられる。また、直近では、グラフ畳み込みニューラルネットワーク（Graph Convolutional Neural Network, GCNN）を用いて、 テキストと単語の難易度を半教師ありで推定する手法が提案されている [10]。

ランキング学習は、文献 [11], [18] が詳しい。この論文では、所与の 1 テキストに対してテキストの難しさのラベル

を予測する教師あり識別の問題設定ではなく、テキストの集合を入力として、これらのテキストを難しさの順番に並び替える「教師ありランキング学習」の問題に帰着させている。

英語のリーダビリティ判定の古典的な研究として、テキストの難しさの段階（テキストが用いられている学年など）に対して、テキスト中の単語の平均長などの回帰式を用いた研究がある。Flesch-Kincaid Grade Level (FKGL) [4] や、SMOG grade[19]、Coleman-Liau index [20] などがこれにあたる。日本語については文献 [21] が詳しい。

言語モデルは、直近で発表された論文 [1] で用いられている手法である。所与のテキストに対して、言語モデルのパープレキシティなどを用いた指標を計算する。個人化リーダビリティについては次節にまとめる。

3. 個人化リーダビリティ

応用言語学分野では、読み手となる学習者が所与のテキストを読めるかどうかを判定する研究が盛んである。この問題設定は、応用言語学分野では 1980 年代からある古典的な問題設定である [13], [14]。

この設定では、まず、読み手となる外国語学習者が事前に語彙テスト（単語テスト）を受けているものとする。そして、その語彙テストの結果を用いて、所与のテキスト中の知っている単語（既知語）を推定し、そこから既知語率を計算し、既知語率が閾値を超えた場合に、学習者がテキストを「読める」と判断する [13], [14], [15]。個人化リーダビリティは、簡単に言えば、「テキスト中で知らない単語の比率が多ければ、テキストは読めないはずだ」という直観に基づく手法である。

単純にはこの通りだが、語彙テストから個々の学習者の既知語をどのように推定するのか、また、既知語率が閾値を超えた場合にテキストが「読める」と判定する事の妥当性の 2 点について、詳述する。

3.1 既知語判定

語彙テストの結果から既知語を推定する点については、理想的には、テキスト中に現れそうなその言語の全ての語種について、学習者が知っているかどうか、学習者をテストする事が望ましいが、これには学習者に膨大な負担がかかり、非現実的である。現実的な方法として、高々数百語程度の語彙テストを、数十分ほど行ってもらい、このテスト結果を利用して、語彙テストに含まれない語を各学習者が知っているかどうかを推定する方法がとられている。例えば、文献 [15] では、100 語からなるテストを考案している。

語彙テストの結果から、語彙テストに含まれない語を各学習者が知っているか推定する手法の 1 つとして、単純に、語彙量 (vocabulary size) を用いた方法が挙げられる [22]。すなわち、全ての学習者が、British National Corpus などの均衡コーパス中の頻度順に語を学習することを仮定し、頻度の高い順に、推定された語彙量番目までの語は全て知っており、それより頻度の低い語については全て知らないと推定することで、既知語判定を行っている。この既知語判定問題については、機械学習の観点からは、語彙テストの結果を訓練データとして、語と学習者が与えられたときに学習者が語を知っているか否かを判定する、単純な二値識別の問題として定式化できる [23], [24]。この 2 値識別の問題に対して、半教師あり学習や能動学習を用いて精度向上した研究が文献 [25] である。また、既知語判定問題の標準的なデータセットについては、筆者が以前作成している [26]。

既知語判定問題は、テキスト中の知らない単語を発見する Personalized Complex Word Identification タスクの一種ともみなせ、テキスト簡単化の個人化などにも応用されている [27]。

3.2 既知語の閾値

学習者がテキストを「読める」既知語率の閾値については、95%または 98%の値が用いられることが多い。英語の既知語率と、テキストが「読める」閾値の関係性の検証については、文献 [28] が代表的である。具体的には、イスラエルの大学入試問題の英語の読解問題で、読み手が合格水準に達している場合に、その読解問題のテキストが「読める」と定義している。

また、既知語率の閾値については、既知語判定問題の識別器が返す、「ある語が既知語である確率」を用いて、所与のテキストの「既知語率の確率分布」を計算し、既知語率の閾値の解釈性を保ったまま性能向上させる手法を、著者は過去に提案している [29]。

3.3 問題設定の違い

このように、個人化リーダビリティは、自然言語処理の典型的なリーダビリティ判定の評価用データセットとは、「リーダビリティ」の信頼性をどこに依拠するかの点で異なっている。自然言語処理の典型的なリーダビリティ判定の評価用データセットは、前述の OneStopEnglish コーパス [30] がそうであったように、基本的には語学教師などで構成される、テキストに対して正解ラベルを付与した「アノテータ」に依拠したリーダビリティである。つまり、「リーダビリティ」と言いながらも、実際に語学学習者がテキストを「読める」かどうかについては直接測定しておらず、その点はアノテータとなる語学教師の判断に依拠している訳である。

一方、個人化リーダビリティは、前述のように、学習者がテキストを「読める」か否かについて、読解問題を通じた検証に基づいているため、学習者がテキストを「読める」

かどうかを直接的に計測して検証されてはいる。ただし、語彙テストからリーダビリティの判定に至るまでに、学習者の既知語の推定と、学習者の既知語率と学習者がテキストを「読める」か否かの推定の 2 つの推定が入っている。このように、複数の不確実な推定のプロセスが入っているにも関わらず、応用言語学分野で個人化リーダビリティが広く使われている理由は、おそらく、既知語率が解釈しやすい概念であること、また、既知語率の閾値が比較的狭い範囲（95%〜98%）で判定できることが服須の研究で示されていることが、貢献していると思われる。その背後には、「テキスト中で知らずに意味を推測しながら読める単語の量には認知的な限界があり、その限界はテキストによって大きく変わらないだろう」という直観があるものと思われる。

3.4 個人化リーダビリティを用いた教師なし自動リーダビリティ判定

本節では、前節で説明した個人化リーダビリティ判定器を用いて、自然言語処理分野で一般的な自動リーダビリティ判定器を作成する手法について詳述する。個人化リーダビリティでは、まず、個々の外国語学習者が知っている語彙を推定する必要がある。これには、100 単語程度の語彙テスト [15] の結果を分析し、この 100 単語以外の単語について、学習者が各単語を知っているかどうかを判定する手法が用いられる。このようにして推定された学習者が知っている語彙から、語彙テストを受けた学習者がテキストを読めるかどうかを判定する [22]。この際には、学習者がテキスト中の 95%〜98% 程度の単語を知っていれば学習者がテキストを読めるとする応用言語学上の知見を用いることが行われている。前述のように、学習者が事前に語彙テストを受けなければならないという設定のためか、応用言語学分野の外では、この手法はあまり用いられていない。

[26] では、100 問の語彙テストについて、クラウドソーシング上で集めた 100 人の被験者の回答が収められている。この語彙テストデータセットは、もちろん、リーダビリティを判定するテキストとは全く関係のないものである。この中で、最も標準的な語彙力を持つ学習者にとっての個人化リーダビリティ判定を、一般的なリーダビリティとして算出する。

語彙テストの分析には項目反応理論 [31] の考え方を応用したモデリングを用いる。これは、語彙テストのようなテストの各設問に対して、被験者が正答/誤答したという結果のデータセットから、被験者の能力値と各設問の難しさを同時に推定する心理モデルである。これは、機械学習の用語を用いれば、本質的には単純な 2 値ロジスティック回帰モデルと同等である。

$\mathcal{V}$ を語彙の集合とし、$\mathcal{L}$ を学習者の集合とする。$z_{v,l} \in \{0,1\}$ を、学習者 $l \in \mathcal{L}$ が語 $v \in \mathcal{V}$ に正答したかどうかとする。$z_{v,l} = 1$ であれば、正答、$z_{v,l} = 0$ であれば誤答とする。$z_{v,l} = 1$ であることは、学習者 l が単語 v を知っていることを示唆する。

次に、$\{z_{v,l}\}$ を訓練データとして、次のモデルを学習する。

$$p(z = 1|v, l) = \text{sigmoid}(a_l - d_v) \qquad (1)$$

(1) で、a_l は学習者 l の能力パラメタ、d_v は単語 v の難しさパラメタである。また、sigmoid はロジスティックシグモイド関数であり、$\text{sigmoid}(x) = \frac{1}{1+\exp(-x)}$ で定義される。

ロジスティックシグモイド関数は、ニューラル識別モデルで用いられる softmax 関数の 2 値版であり、$(0, 1)$ の範囲での単調増加関数である。$\text{sigmoid}(0) = \frac{1}{1+1} = \frac{1}{2}$ であるので、学習者の能力パラメタ a_l が単語の難しさパラメタ d_v より大きければ、学習者が単語を知っている確率が 1/2 を超える。このように、学習者の能力と単語の難しさを同じ尺度で比較できるのが、項目反応理論の大きな特徴の 1 つである。

(1) だけでは、語彙テスト結果データセット中で設問に現れる単語の難易度しか d_v として計算する事ができない。語彙テスト結果データセットで設問とされている以外の単語について d_v を知るためには、d_v をコーパス中の単語頻度などの特徴量から求めればよい。具体的には、次のようにして構成した。

$$d_v = -\sum_{k=1}^{K} w_k \log(\text{freq}_k(v) + 1) \qquad (2)$$

(2) で、K は使用するコーパスの数、k は k 種類目のコーパスを表し、$\text{freq}_k(v)$ は、k 種類目のコーパス中での単語 v の頻度である。また、w_k は、このコーパスに対する重みパラメタである。(2) で全体に負号がついているのは、一般に、コーパス中の単語頻度が大きくなるほど単語は簡単になるので、単語の難しさとは逆の尺度であるためである。

パラメタ推定に必要な情報をまとめよう。$\{z_{v,l}\}$ と、コーパスの単語頻度 $\text{freq}_k(v)$ が与えられれば、学習者 l の能力値パラメタ a_l とコーパス k の重みパラメタ w_k が推定できる。(1) と (2) をまとめると、sigmoid 関数内がパラメタに対して線形であるため（つまり、2 種類のパラメタの積から構成される項が存在しないため）、これはロジスティック回帰を使って表現する事ができることがわかる。実際、実験では、Python の機械学習パッケージである scikit-learn [*1]を用いた。scikit-learn は、内部的にはロジスティック回帰の高速実装として有名な LIBLINEAR [*2]を呼び出している。

このようにしてパラメタを求めた後、所与のテキスト $\mathcal{T}$ に対するリーダビリティを判定する。簡単には、最も a_l が標準的な学習者 l_{avg} を 1 人選び、この学習者がこのテキスト中の各単語を知っている確率をつぎのように求めればよ

[*1] https://scikit-learn.org/stable/
[*2] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

い。ここで、$v \in \mathcal{T}$ は、テキスト中の単語 v を表し、$\mathcal{V}(\mathcal{T})$ はテキスト中で現れる語彙の集合を表す。例えば、テキスト中で 3 種の単語が出現したら、$v \in \mathcal{T}$ では、頻度の回数分、各単語の確率値をかけ合わせるのに対して、$\mathcal{V}(\mathcal{T})$ では、単純に語種の数だけでかけ合わせる。

$$score_{\mathrm{sum}}(\mathcal{T}) = -\log\left(\prod_{v\in\mathcal{T}} p(z=1|v,l_{\mathrm{avg}})\right) \quad (3)$$

$$score_{\mathrm{vsum}}(\mathcal{T}) = -\log\left(\prod_{v\in\mathcal{V}(\mathcal{T})} p(z=1|v,l_{\mathrm{avg}})\right) \quad (4)$$

$$score_{\mathrm{vavg}}(\mathcal{T}) = -\frac{1}{|\mathcal{V}(\mathcal{T})|}\log\left(\prod_{v\in\mathcal{V}(\mathcal{T})} p(z=1|v,l_{\mathrm{avg}})\right) \quad (5)$$

また、$p(z=1|v,l_{\mathrm{avg}})$ が計算できれば、上式に変わり、テキスト中の 95% の単語知っている確率を求め、これをスコアにする方法もある [29]。パラメタ推定の際には、リーダビリティ評価用データセットのテキストの難しさラベルは一切使用しないため、この手法は「教師なし」に分類される。

最後に、以上で述べた個人化リーダビリティ判定においては、語彙テスト結果から、単語の難しさパラメタを求める部分 (2) が本質であることを説明する。説明のため、最も平均的な能力の学習者を 1 人定めて l_{avg} としたが、(1) では sigmoid 関数は単調増加関数であること、a_l は単純に d_v に足されていることから、実は、どの学習者を選んでも、a_l を固定した時点で、学習者が単語を知っている確率 $p(z=1|v,l$ に寄与するのは d_v だけである。従って、上記の方法は、単語テスト結果を用いて、単語テスト結果とよく相関するような単語の難易度を、コーパス中の単語の頻度 $\mathrm{freq}_k(v)$ から作り出す手法であると捉えられる。

4. 実験結果と考察

データセットには、第二言語学習者を対象にしたデータセットとして比較的最近に報告されたものであることから、OneStopEnglish データセットを用いた [7]。このデータセットは、Guardian 誌の記事を語学教師が Elementary, Intermediate, Advanced の 3 種類に書き換えたものである。各レベルには 189 件のテキストがあり、全体では 567 件である。教師あり手法とも比較するため、これを、339 件の訓練データ、114 件の開発データ、114 件のテストデータに分割し、最後のテストデータを用いて性能検証を行う。比較手法は、下記の通りである。

古典的な自動リーダビリティ判定式については、Python の **readability** パッケージを用いて実装した [*3]。このパッケージには、英語のリーダビリティ判定式として、Flesch-Kincaid Grade Level[4], ARI, Coleman-Liau Index, Flesch Reading Ease, Gunning Fog Index, LIX, SMOG Index [19], Dale-Chall Index が実装されているのでこれを用いた。紙面の都合上、全ての数式をここに表記する事はしない。具体的な式については、脚注に記した **readability** パッケージのプロジェクトページに記載がある。これらの手法は、全てリーダビリティのラベルを用いないので、「教師なし」に分類される。

次に [1] において提案されている、ニューラル言語モデルを用いた教師なし自動リーダビリティ判定について説明する。ニューラル言語モデルについては、事前学習モデル **bert-large-cased-whole-word-masking** を Huggingface の事前学習モデル一覧より取得し、これを用いて計測した各テキストのパープレキシティの平均値をリーダビリティとしたのが **BERTLMavg** である。[1] では BERT を用いた言語モデルとしては **bert-base-uncased** を事前学習モデルに使用したものが用いられているが、**bert-large-cased-whole-word-masking** はこれより大きなモデルである。テキストの文分割については、**nltk** パッケージ[*4]の **sent_tokenize** 関数を用いた。

[1] では、BERT の言語モデルを用いた手法はよい性能をあげられていないが、そのほかの手法は公開されている事前学習モデルを用いておらず、再実装が難しい。そこで、[1] の OneStopEnglish データセットでの最高性能を達成している **TCN RSRS-simple** の結果を、実験結果の表に加えた。ただし、[1] で用いたテストデータが入手できなかったため、この手法は直接の比較が可能ではないため、表中では (*) を用いてそのことを明記した。**TCN RSRS-simple** は、単純に言えば、Temporal Convolutional Network (TCN) を Simplified Wikipedia コーパス上で事前学習させ、さらに、パープレキシティにかわり、Ranked Sentence Readability Score (RSRS) という [1] が独自に定義した指標を用いて判定するものである。**TCN RSRS-simple** のさらなる詳細については [1] を参照されたい。

最後に、**提案手法**が本研究の提案手法である。これは、語彙テストデータセット [26] を用いて、前述のパラメタ推定を行い、(3) を用いて学習者が各単語を知っている確率を自動リーダビリティ判定に用いたものである。コーパスからの単語頻度の特徴量としては、英語教育上広く使われていることから、British National Corpus [*5]と Corpus of Contemporary American English (COCA) [*6]を用いた。さらに、単純に、これらのコーパス頻度を表す特徴量を **BNC**、**COCA** として結果表中に掲載した。

表 2 に結果を示す。[1] では、スコアとリーダビリティ評価用データセットのラベルとの相関として Pearson's ρ しか用いていないが、これは、スコアの線形性が低いとスコアが下がってしまう事から、順位相関係数として Speraman's

[*3] https://pypi.org/project/readability/

[*4] nltk.org

[*5] https://www.english-corpora.org/bnc/

[*6] https://www.english-corpora.org/coca/

表 2 OneStopEnglish データセットでの実験結果・考察。式 (3) と式 (4) が提案手法。

教師あり/なし	手法	Spearman's ρ	Kendall's τ-b	Pearson's ρ
教師なし	Flesch-Kincaid	0.324	0.253	0.359
	ARI	0.317	0.248	0.351
	Coleman-Liau	0.373	0.295	0.372
	FleschReadingEase	-0.387	-0.301	-0.426
	GunningFogIndex	0.331	0.257	0.362
	LIX	0.348	0.273	0.383
	SMOGIndex	0.456	0.360	0.479
	RIX	0.437	0.340	0.462
	DaleChallIndex	0.495	0.387	0.506
	TCN RSRS-simple	-	-	0.615(*)
	BERTLMavg	-0.220	-0.173	-0.040
	BNC	-0.012	-0.009	-0.006
	COCA	0.018	0.016	0.039
	式 (3)　（提案手法）	0.730	0.592	0.715
	exp(式 (3))	0.730	0.592	0.260
	式 (4)（提案手法）	**0.760**	**0.617**	**0.754**
	式 (5)	0.581	0.454	0.589
教師あり	spvBERT_half	0.751	0.729	0.747
	spvBERT	0.866	0.856	0.864

ρ、Kendall's τ を用いた。また、相関係数の実装には **scipy** を用いた。

リーダビリティ評価用データセットでは、一般に、同じ難しさレベルのテキストが多くあるため、同順を多く含むデータセットになっており、同順補正の方法によってスコアが大きく影響を受ける。各データを順位に変換したものの相関係数として定義される Spearman's ρ については、**midrank** と呼ばれる同順補正法を用いた [32]。ケンドールの τ については、一般に使われている同順補正は τ-b であり、単に Kendall's τ（ケンドールの順位相関係数）と言った場合、こちらが使用されることが多い。片側が 5 件法による尺度でもう片側が 10 件法による尺度である場合など、尺度の細かさに違いがある場合、τ-c という補正を用いた方が良いという報告がある[*7]。しかし、Spearman's ρ で midrank を同順補正法に用いた場合、この様な報告は特にないこと、また、τ-c は scipy のバージョンによっては実装されていないことから、今回は τ-b のみを参考のために表示した。

表 2 の最も左側には、教師なし、教師ありの分類を示した。

最初に、全ての「教師なし」の手法において、提案手法のうち、式 (4) が全ての尺度で最も良い性能を示した。提案手法は 0.760 と、後述の教師ありの設定で訓練データが少ない場合である spvBERT_half を超える順位相関を達成した。式 (4) は、頻度を考慮して足し合わせる式 (3) よりもよい性能を達成しており、テキスト中の語種数に重要な情報が格納されていることが示唆される。このことは、(5) によって、テキスト中の語種数で平均を取り、テキスト中の語種数の情報を反映しないと、著しく精度が下がることからもわかる。

次に、テキスト中の語のうち、難しい語（学習者が知っている確率が低い語）がリーダビリティ判定において重要なのか、簡単な語が重要なのかを確認した。(4) で、テキスト中の難しい語種上位 30 語を削ると、Spearman's ρ が 0.769 まで向上した。一方、簡単な語種上位 30 語を削った場合、0.761 であった。この結果は、テキスト中の難しい語種には、実際に難しい語の他にも、“Redmond” や “Stockholm” といった固有名詞も含まれてしまっており、難しい語種を削ることによって、こうしたノイズが削減されるためであると考えられる。このように、式 (4) が式 (3) を性能で上回っているものの、(3) も（既存手法のテストデータが入手できなかったため直接比較は難しいものの）既存手法を超える性能を達成しているので、両式を提案手法とした。次に、提案手法以外の手法との比較を具体的にみていく。

Pearson's ρ がスコアの線形性に影響される度合いを調べるために、式 (3) のスコアに exp をかませ、スコア s に対して $\exp(s)$ をスコアとしたものを exp(式 (3)) として表 2 に示した。exp は単調増加関数であるため、スコアの順位には影響しないので、順位相関の尺度は元の提案手法の性能と変わらないが、Pearson's ρ では、0.260 と著しく低い値が出ている。このため、スコアの線形性が担保されない状況では、Pearson's ρ を評価尺度に使うことは望ましくないことがわかる。

BERTLMavg は [1] よりも大きな事前学習モデルを用

[*7] https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

表 3 spvBERT の混同行列

		予測		
		Elementary	Intermediate	Advanced
Gold Labels	Elementary	39	0	2
	Intermediate	1	34	2
	Advanced	2	2	32

いてパープレキシティを計測したが、良い結果を示さなかった。これは、パープレキシティが第二言語学習者向けのリーダビリティの尺度として適していないことを示唆する。

TCN RSRS-simple は [1] における OneStopEnglish データセット上の最高性能を達成した手法である。[1] においては、性能比較に Pearson's ρ のみが用いられているため、この値だけを表示した。ただし、彼らは同じ OneStopEnglish データセットを用いてはいるが、性能値を算出するために具体的にどのデータをテストデータに用いたのかが公開されていないため、直接の比較は難しく、(*) でこのことを明示した。直接の比較は難しいものの、提案手法は、**TCN RSRS-simple** よりもよい性能を達成できていることがわかる。

また、おもしろいことに、**BNC** と **COCA** の単語頻度については、英語教育の分野では単語の難しさを測る良い指標とされているものの、これら単独ではリーダビリティ評価用データセットのラベルと良い相関が得られなかった。一方、**提案手法**では、前述のように、これらの単語頻度特徴量を (2) を用いて組合せ、語彙テストデータセットに沿う単語難易度を求めている。このことから、複数のコーパスからの単語頻度を組み合わせて、「第二言語学習者にとっての単語の難しさ」をきちんと語彙テストデータから計測することが、自動リーダビリティ判定に重要であることが示唆される。

教師あり学習の手法の結果を示す。**spvBERT** は BertForSequenceClassification 関数を用いてリーダビリティラベルを用いて学習した結果であり、**spvBERT_half** は、訓練データを半分にして同じ学習をした場合である。モデルとしては、前述の **bert-large-cased-whole-word-masking** を用いた。教師データを用いることにより、spvBERT は教師なしである提案手法より高い性能を達成できている。

最後に、表 2 からの教育の観点からの説明性について考察する。spvBERT は、教師あり学習であり、テキスト全体の文脈を見て判別する手法である。一方、提案手法は、教師なし学習ではあるが、単語の難しさについては語彙テストデータセットを用いて正確に求める手法である。提案手法は、単語の難しさについては正確に求めるものの、文脈については見ていない。従って、**spvBERT の性能値と、提案手法の性能値の差が、リーダビリティ判定を平均的な単語難易度だけではなく、文脈を見て行う事による性能向上であると考えることができる。**

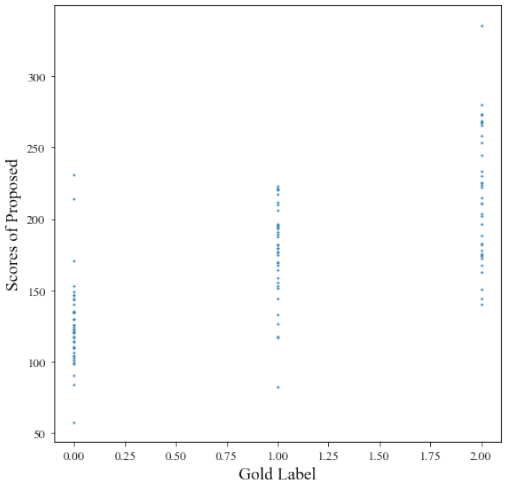

図 1 提案手法（**式** (3), Proposed）の散布図

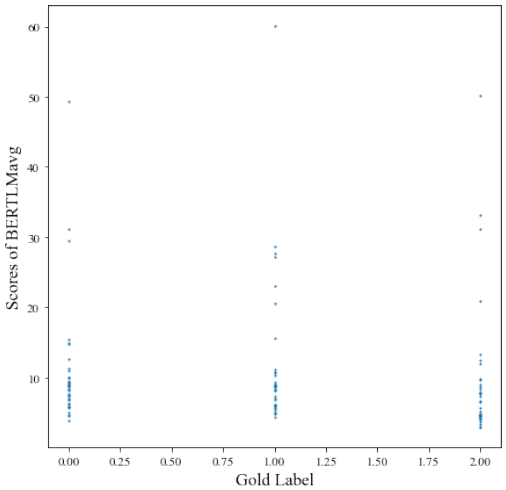

図 2 **BERTLMavg** の散布図

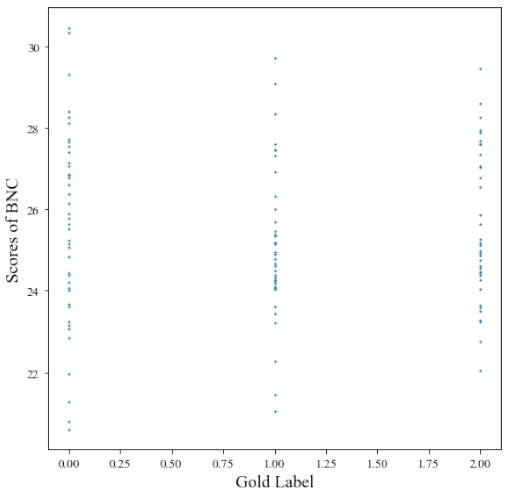

図 3 **BNC** の散布図

表 4 spvBERT_half の混同行列

		予測		
		Elementary	Intermediate	Advanced
Gold Labels	Elementary	38	0	3
	Intermediate	4	29	4
	Advanced	4	4	28

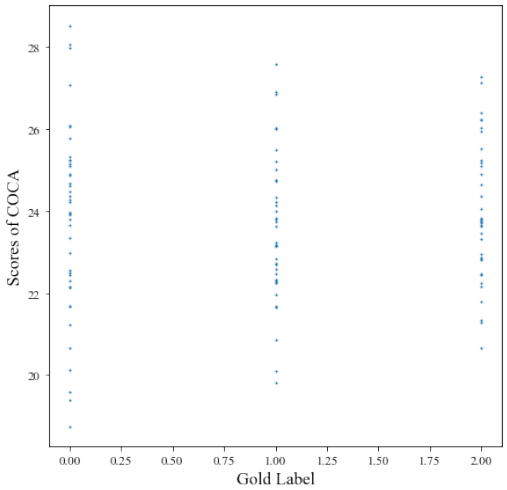

図 4　COCA の散布図

4.1　詳細な分析

図 1, **図 2**, **図 3**, **図 4** に散布図を示す。各散布図中の各点はテストセットの各テキストを表す。横軸は真のラベル（Gold Labels）を表し、0：Elementary（初級）、1：Intermediate（中級）、2：Advanced（上級）である。このように、図 1 のスコアは、真のラベルとよく相関していることがわかる。また、提案手法（式 (3)）は、ラベル 2 とラベル 1 を区別する事は難しいものの、ラベル 1 とラベル 0 をより明確に区別していることがわかる。

図 2 のスコアと真のラベルの間には、若干の負の相関があります。この負の相関は、パープレキシティが低いほど流暢性が高いことを意味するので、その点では理解できる。しかし、点が重なっている部分が多く、そのために相関が低くなっていることが、見て取れる。この結果は、表 2 でも確認できる。また、図 3 と図 4 は、スコアと真のラベルの間に相関がないことがわかる。これらの結果は、表 2 の低い相関値に対応している。

以上は教師なしリーダビリティ判定器の性能比較であった。次に、教師ありリーダビリティ判定器の性能についても詳細な分析を行う。**表 3** と**表 4** には、それぞれ、**spvBERT** と **spvBERT_half** の混同行列を示す。行は真のラベル（Gold Labels）、列は予測値を示している。精度が、それぞれ 0.92、0.83 であることから、学習データを半分に分けると、予測の精度が落ちることが確認できる。また、２つの混同行列から、初級と中級の区別は、中級と上級の区別よりもはっきりしていることが読み取れる。興味深いことに、どちらの表でも、中級と予測されたテキストのうち、初級であるテキストは存在しなかった。また、上級と予測された 2 や 3 のテキストが実際には初級であることもあることから、このような外れ値のテキストがあることも見てとれる。

4.2　メモリと計算速度

人手のリーダビリティ判定のラベルを訓練に使わない教師なしの BERT 言語モデルである **BERTLMavg** は、1,793 MiB の GPU メモリを使用した。当然、GPU を搭載している計算機上でないと動作しない。これに対して、提案手法は単なるロジスティック回帰であるため、GPU メモリは全く必要としないので、GPU を搭載していない機器上でも動作させることが可能である。さらに、提案手法は CPU メモリもあまり消費しない。特徴量として用いているのは、BNC と COCA であるため、CPU メモリは 10MiB 程度であった。速度面では、テストセットの全テキストを分類するのに、**BERTLMavg** は 368 秒かかったのに対し、提案手法は 5.37 秒しかかからなかった。つまり、提案手法は 68.5 倍高速に動作した。

表 5　spvBERT による様々なテキストのリーダビリティ判定結果。数値は比率。

データの種類	Elementary	Intermediate	Advanced
経済ニュース	0.004	0.962	0.034
ACL Anthology	0.037	0.860	0.103
PubMed	0.006	0.639	0.305
W-NUT 正規化前	0.486	0.512	0.002
W-NUT 正規化後	0.495	0.503	0.002

教師なしモデルと比較して、人手のリーダビリティ判定のラベルを訓練に使用する教師あり BERT モデルである **spvBERT** は、約 8 倍も大きい 16GiB の GPU メモリを使用する。参考までに、提案手法の訓練時間と、**spvBERT** の訓練時間を比較する。**spvBERT** はリーダビリティ判定のラベルを用いて訓練するのに 162 秒かかったが、提案手法は語彙テスト結果データを用いて訓練するのに 9.7 秒しかかからなかった。このように、提案手法は、教師なしモデルである **BERTLMavg** よりも、教師ありモデルである **spvBERT** よりもはるかに高速に動作する。

5.　様々なテキストのリーダビリティ

次に、様々な種類のテキストに対してリーダビリティ判定を行い、リーダビリティ判定が定性的に直観にそっていることを確認する。以下、全てのリーダビリティ測定には、**spvBERT** を使用した。

まず、経済ニュース記事のデータセット [33] を用いて、経済ニュースのリーダビリティを求めた。このデータセットは、2006 年〜2007 年ごろの Reuters の経済ニュースを収集したものである。このうち、2007 年の経済ニュース全てのリーダビリティを判定した。次に、科学分野のテキストを入手するため、ACL Anthology (`https://aclanthology.org/`) と PubMed(`https://pubmed.ncbi.nlm.nih.gov/`) から論文概要を 1,000 件ずつランダムに取り出し、リーダビリティ判定を行った。最後に、Twitter などの口語表現に近いテキストのリーダビリティのデータセットとして、W-NUT のテキスト正規化データセットのリーダビリ

ティを判定した（http://noisy-text.github.io/2021/multi-lexnorm.html）。

結果を**表 5** にまとめる。まず、W-NUT の Twitter のテキストのデータは、他のテキストより Elementary の比率が多く、簡単である事、また、正規化によって簡単なテキストの割合が増えることが分かる。これは、Twitter のテキストは口語表現に近く、また短文が多いためであると思われる。次に、経済ニュースは大部分のテキストが Intermediate であり、これは人為的にそうなるように書かれていることが示唆される。最後に、ACL Anthology や PubMed などの科学分野のテキストは、今回比較した他の種類のテキストより難しい事、また、PubMed は特に難しいテキストが多いことが分かる。このように、**spvBERT** は、定性的にも直観に沿っていることが示唆された。

6. 議論

自然言語処理においては、本研究で扱ったアプローチは、第二言語学習者にとって難しい単語を発見する複雑単語同定（Complex Word Identification, CWI）のタスクと密接な関係がある [34]。複雑単語同定と、個人化リーダビリティの研究については [29] が詳しい。特に、個々の学習者にとって難しい単語を発見するタスクは、個人化複雑単語同定（Personalized CWI）と呼ばれる [25], [35]。個人化複雑単語同定は、下流に様々な自然言語処理上の応用があり、個人化テキスト簡単化 [17]、言語学習者のためのテキスト推薦 [24], [36]、クラウドソーシング翻訳の翻訳者選択 [37] などに用いられている。

7. まとめ

本研究では、第二言語学習者を対象とした教師なしの自動リーダビリティ判定タスクに注目した。学習者の第二言語の知識を効率的に利用するために、学習者の語彙テストの結果をリーダビリティ判定に利用するシンプルなアプローチを提案した．実験の結果、我々の手法は大規模な BERT ベースのニューラル言語モデルよりも優れていることがわかった。提案モデルは、語彙テスト結果のデータセットで学習されたロジスティック回帰のみを使用しているため、より軽量であり、メモリ使用量も少なく、スマートフォンなどの計算リソースの少ないマシンでも動作する。

また、教師ありリーダビリティ判定手法として**spvBERT** を様々なテキストに適用し、その判定結果が定性的に直観に沿っていることを確認した。

今後の課題としては、単語の難易度を超えて単語の難易度に寄与する重要な情報を特定するために、**SpvBERT** と提案手法の詳細な誤り分析を行う事が挙げられる。また、信頼性の高い他の評価データセットを用いた追加実験が挙げられる。特に、今回語彙テスト結果から用いた単語難易度は教師なしリーダビリティ判定器の構築に有用であることが分かったが、この単語難易度の尺度と、CEFR-J Vocabulary Profile（https://github.com/openlanguageprofiles/olp-en-cefrj）など、他の研究で注目されている単語難易度の尺度との関連性などが挙げられる。

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。

参考文献

[1] Martinc, M., Pollak, S. and Robnik-Šikonja, M.: Supervised and Unsupervised Neural Approaches to Text Readability, *Computational Linguistics*, Vol. 47, No. 1, pp. 141–179 (2021).
[2] Dale, E. and Chall, J. S.: A formula for predicting readability: Instructions, *Educational research bulletin*, pp. 37–54 (1948).
[3] Flesch, R.: A new readability yardstick, *Journal of Applied Psychology*, Vol. 32, No. 3, pp. 221–233 (1948).
[4] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, Technical report, Naval Technical Training Command Millington TN Research Branch (1975).
[5] Feng, L., Jansche, M., Huenerfauth, M. and Elhadad, N.: A Comparison of Features for Automatic Readability Assessment, pp. 276–284 (online), available from ⟨https://www.aclweb.org/anthology/C10-2032⟩ (2010).
[6] Xia, M., Kochmar, E. and Briscoe, T.: Text Readability Assessment for Second Language Learners, *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, San Diego, CA, Association for Computational Linguistics, pp. 12–22 (online), DOI: 10.18653/v1/W16-0502 (2016).
[7] Vajjala, S. and Lučić, I.: OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 297–304 (online), DOI: 10.18653/v1/W18-0535 (2018).
[8] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. of NAACL*, Minneapolis, Minnesota, pp. 4171–4186 (2019).
[9] Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M.: Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, Rochester, New York, Association for Computational Linguistics, pp. 460–467 (online), available from ⟨https://www.aclweb.org/anthology/N07-1058⟩ (2007).
[10] Fujinuma, Y. and Hagiwara, M.: Semi-Supervised Joint Estimation of Word and Document Readability, *arXiv:2104.13103 [cs]*, (online), available from ⟨http://arxiv.org/abs/2104.13103⟩ (2021). arXiv:

2104.13103.
[11] Tanaka-Ishii, K., Tezuka, S. and Terada, H.: Sorting Texts by Readability, *Computational Linguistics*, Vol. 36, No. 2, pp. 203–227 (online), DOI: 10.1162/coli.09-036-R2-08-050 (2010).
[12] Flesch, J.: Flesch-Kincaid readability formula (1965).
[13] Nation, P.: *Teaching and Learning Vocabulary*, Heinle and Heinle, Boston, MA (1990).
[14] Laufer, B.: What percentage of text-lexis is essential for comprehension, *Special language: From humans thinking to thinking machines*, Vol. 316323 (1989).
[15] Beglar, D. and Nation, P.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).
[16] Ehara, Y., Sato, I., Oiwa, H. and Nakagawa, H.: Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty, *Journal of Information Processing*, Vol. 26, pp. 267–275 (online), DOI: 10.2197/ipsjjip.26.267 (2018).
[17] Lee, J. and Yeung, C. Y.: Personalized Substitution Ranking for Lexical Simplification, *Proceedings of the 12th International Conference on Natural Language Generation*, Tokyo, Japan, Association for Computational Linguistics, pp. 258–267 (online), DOI: 10.18653/v1/W19-8634 (2019).
[18] 佐藤理史：均衡コーパスを規範とするテキスト難易度測定，情報処理学会論文誌，Vol. 52, No. 4, pp. 1777–1789 (2011).
[19] Mc Laughlin, G. H.: SMOG grading-a new readability formula, *Journal of reading*, Vol. 12, No. 8, pp. 639–646 (1969).
[20] Coleman, M. and Liau, T. L.: A computer readability formula designed for machine scoring., *Journal of Applied Psychology*, Vol. 60, No. 2, p. 283 (1975).
[21] Hasebe, Y. and Lee, J.-H.: Introducing a readability evaluation system for Japanese language education, *Proceedings of the 6th international conference on computer assisted systems for teaching & learning Japanese*, pp. 19–22 (2015).
[22] Nation, I.: How Large a Vocabulary is Needed For Reading and Listening?, *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82 (2006).
[23] Ehara, Y., Shimizu, N., Ninomiya, T. and Nakagawa, H.: Personalized Reading Support for Second-language Web Documents by Collective Intelligence, *Proc. of IUI*, IUI '10, ACM, pp. 51–60 (online), available from ⟨http://doi.acm.org/10.1145/1719970.1719978⟩ (2010). event-place: Hong Kong, China.
[24] Ehara, Y., Shimizu, N., Ninomiya, T. and Nakagawa, H.: Personalized Reading Support for Second-language Web Documents, *ACM Trans. Intell. Syst. Technol.*, Vol. 4, No. 2, pp. 31:1–31:19 (online), DOI: 10.1145/2438653.2438666 (2013).
[25] Ehara, Y., Miyao, Y., Oiwa, H., Sato, I. and Nakagawa, H.: Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning, *Proc. of EMNLP*, pp. 1374–1384 (online), DOI: 10.3115/v1/D14-1143 (2014).
[26] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).
[27] Lee, J. and Yeung, C. Y.: Personalizing Lexical Simplification, *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, Association for Computational Linguistics, pp. 224–232 (online), available from ⟨https://www.aclweb.org/anthology/C18-1019⟩ (2018).
[28] Laufer, B. and Ravenhorst-Kalovski, G. C.: Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension, *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30 (online), available from ⟨https://eric.ed.gov/?id=EJ887873⟩ (2010).
[29] Ehara, Y.: Uncertainty-Aware Personalized Readability Assessments for Second Language Learners, *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1909–1916 (online), DOI: 10.1109/ICMLA.2019.00307 (2019).
[30] Vajjala, S. and Rama, T.: Experiments with Universal CEFR Classification, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, Louisiana, Association for Computational Linguistics, pp. 147–153 (online), DOI: 10.18653/v1/W18-0515 (2018).
[31] Baker, F. B.: *Item Response Theory : Parameter Estimation Techniques, Second Edition*, CRC Press (2004).
[32] 神嶌敏弘：順序の距離と確率モデル，人工知能学会第二種研究会資料， Vol. 2009, No. DMSM-A902, p. 07（オンライン），DOI: 10.11517/jsaisigtwo.2009.DMSM-A902$_0$7(2009).
[33] Ding, X., Zhang, Y., Liu, T. and Duan, J.: Using Structured Events to Predict Stock Price Movement: An Empirical Investigation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Association for Computational Linguistics, pp. 1415–1425 (online), DOI: 10.3115/v1/D14-1148 (2014).
[34] Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A. and Zampieri, M.: A Report on the Complex Word Identification Shared Task 2018, *arXiv:1804.09132 [cs]*, (online), available from ⟨http://arxiv.org/abs/1804.09132⟩ (2018). arXiv: 1804.09132.
[35] Ehara, Y., Sato, I., Oiwa, H. and Nakagawa, H.: Mining Words in the Minds of Second Language Learners: Learner-Specific Word Difficulty, *Proceedings of COLING 2012*, Mumbai, India, The COLING 2012 Organizing Committee, pp. 799–814 (2012).
[36] Yeung, C. Y. and Lee, J.: Personalized Text Retrieval for Learners of Chinese as a Foreign Language, *Proc. of COLING*, pp. 3448–3455 (2018).
[37] Ehara, Y., Baba, Y., Utiyama, M. and Sumita, E.: Assessing Translation Ability through Vocabulary Ability Assessment, *Proc. of IJCAI* (2016).