Vol.2021-CH-127 No.6
2021/8/28

応用言語学と自然言語処理のリーダビリティ判定に対するアプローチの違い

江原　遥[1,a)]

概要：リーダビリティ判定は、テキストを入力として、そのテキストの読みやすさを自動的に判定する手法である。応用言語学の分野においては、語彙テストの結果をもとに個々の学習者が所与のテキストを読解可能か判定する個人化リーダビリティ判定の研究が活発に研究されている。また、自然言語処理の分野においても、リーダビリティ判定の研究は盛んにおこなわれており、近年は、特に第二言語学習者向けの性能評価用のデータセットも整備されてきている。しかし、この 2 つのアプローチの研究は、双方のコミュニティで活発かつ継続的に研究されているにもかかわらず、相互の引用がほとんどない。本発表では、まず、両系統の手法を機械学習の観点から分類し、機械学習の観点からは両系統の手法の性能比較が可能であることを示す。そして、相互の引用が少ない理由について、両アプローチでの問題設定・目的の違いから考察する。そのうえで、応用言語学の手法を用いた自動リーダビリティ判定器を作成したところ、ニューラル言語モデルなどを駆使した既存手法より高精度かつ高速であったという実験結果について報告する。

1. はじめに

応用言語学の分野においては、語彙テストの結果をもとに個々の学習者が所与のテキストを読解可能か判定する個人化リーダビリティ判定の研究が活発に研究されている [20], [26]。また、自然言語処理、特に教育のための自然言語処理において、テキストのリーダビリティ判定は学習用テキスト推薦などに応用を持つ重要な課題である [12], [29], [31]。にもかかわらず、両系統の研究は積極的に相互に引用されているとは言えない。

本稿では、両系統の手法を分析する。まず、自然言語処理側の手法が「語学教師等、テキストをラベル付けしたアノテータ」に信を置いた手法であることを説明する（節 2）。これは、語学教師を人間の専門家として考え、その専門家の知識を信用するという、知能情報学の方法論の一例になっている。一方で、「リーダビリティ判定」と言いながら、学習者がテキストを読めるかどうかについて学習者から直接の情報を得てはおらず、専門家である語学教師を通じて情報を得る形になっている（節 3）。一方、その語学教師の専門家である応用言語学側では、語学教師自身の判断の根拠を客観的に議論するために、「外国語学習者から直接得た一次情報」に信を置いた手法になっていることを説明する。これは情報源の選択としては適切だが、学習者から直接データを取ることは学習者への負担が大きい。例えば、リーダビリティであれば、実際に個々の学習者に何らかのテキストを読んでもらい読解問題を解いてもらうことが考えられるが、これには時間がかかる。この点は応用言語学分野でも認識されており、対案として、個々の学習者に対しては 30 分〜1 時間程度で回答できる簡便な語彙テストを受けてもらい、その結果からリーダビリティ判定を行う「個人化リーダビリティ」の手法がとられてきた。しかし、ここから実際に自動リーダビリティ判定に至るためには、「語彙テスト結果から学習者の既知語を判定する」、「学習者の既知語から学習者がテキストを読めるかどうか判定する」といった検証を踏まなければならず、この 2 点がどの程度の学習者にどれほど当てはまるのかについては、検証が必要である。

実際には個々の学習者に対して語彙テストと十分な量の読解テストの双方を解いてもらったデータセットがあればよいが、こうしたデータセットで受け入れられているデータセットは、どちらの分野においても、本原稿の執筆時点では、知る限りない[*1]。応用言語学側の個人化リーダビリティ判定の設定では、人間の専門家に判断に基づくテキストの難しさのデータを評価しておらず、リーダビリティ判定器の性能を評価する形式にはなっていないため、学習者

[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

[*1] 読解テストの量が十分とは言えないが、著者はこの形式のデータセットをクラウドソーシング上で作成してはいる [7]

からの一次情報を用いていないことを承知の上で、リーダビリティ判定器の性能を評価するには、自然言語処理側の判定評価用データセットを用いて性能評価する方法が考えられる。

自然言語処理分野で行われているテキストのリーダビリティ判定の性能評価用データセットのフォーマットは、共通している [12], [29], [31]。テキストの難易度をあらかじめ何段階で表すか決めておき、各テキストについて、そのテキストの難しさの段階を示す**正解ラベル**が人手で付与されている。例えば、第 2 言語としての英語学習者 (English as a Second Language, ESL) 向けの性能評価用データセットのうち代表的なものの 1 つである OneStopEnglish コーパス [29] では、英語教師が各テキストについて、Elementary, Intermediate, Advanced の 3 段階でラベル付けが行われている。

こうした評価用データセットの正解ラベルを訓練・評価の際にどのように用いるかによって、既存手法は**表 1** のように分類できる。教師あり/なしは、リーダビリティの判定器を作成する際に、正解ラベルを指標する必要があるかどうかを示している。「可能な性能評価尺度」は、そのリーダビリティ判定器の判定結果を正解ラベルと照らし合わせて性能を評価する際に、どのような尺度を用いることが可能かを示している。

本論文の貢献は、1) 外国語のテキストのリーダビリティ判定の各手法の問題設定を表 1 にまとめたことである。また、2) 自然言語処理をはじめとした知能情報学分野ではあまり知られていないものの応用言語学分野では多くの研究がある「個人化リーダビリティ」が、この分類では順位相関を用いて他の手法と比較可能であることを示したことも貢献である。特に、2) の一連の手法については、自然言語処理のリーダビリティ判定の既存のサーベイ論文 [4] では、まったく触れられていない。本稿は、この点を補足し、分野間の相互理解に貢献したい。以降の節では、各手法について概説し、なぜ表 1 のようにまとめられるのかについて説明している。

さらに、3) 個人化リーダビリティを用いた自動リーダビリティ判定手法を提案した。この手法は、BERT などの大規模な言語モデルから大きなメモリや計算時間を必要とする大規模なニューラル言語モデルを使わずに、教師なし設定で高精度にリーダビリティ判定が行えることをしめしたことである。

2. 自動リーダビリティ判定

本節では表 1 のうち、言語モデルスコアまでの手法を紹介する。これらの手法は、アノテータが付与したリーダビリティのラベルの扱い方に関する方法論であり、ラベルに信を置いた手法になっていることが分かる。**ラベル予測** これらの手法は、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。具体的には、リーダビリティ判定の評価用データセットの一部を訓練データとして切り出し、テキストとラベルのペアを訓練データとして用いて、識別器を訓練する。そして、訓練済みの識別器をリーダビリティのラベルの判定器として用いて、テキストが 1 つ与えられたときに、テキストに紐づいたラベルを予測することを目的とする。すなわち、判定器は直接、ラベルを出力する。代表的な研究としては、文献 [12], [16], [29], [31] など多くの既存研究が挙げられる。近年では、古典的な特徴量抽出と識別器を用いる方法ではなく、事前学習された Bidirectional Encoder Representations from Transformers (BERT) [5] のような深層学習モデルを用いて、転移学習を用いて分類する手法 [17] もあるが、「ラベル付きテキストを訓練データに用いる」という点、「テキストに付与されたラベルを当てる」という点は共通しているため、これらも教師あり識別問題に帰着する。

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	[12], [14], [16], [29], [31]	識別精度、順位相関
	ランキング学習	[28]	順位相関
教師なし	回帰スコア	[13]	
	言語モデルスコア	[24]	
	個人化リーダビリティ	[2], [9], [19], [23], [27]	

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。

教師あり学習の設定では、判定器の出力はリーダビリティのラベルであるため、これを正解ラベルと照らし合わせることにより、単純な識別精度が性能評価指標として用いられることが多い。また、テストデータ用のテキスト集合に対して判定器の出力したラベル集合と、正解ラベル集合の間で、適切に同順補正を行った順位相関係数を用いても性能評価を行うことができる。

ランキング学習 ランキング学習については、文献 [28] が先鞭をつけている。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。この問題設定では、テキストのペアに対して、データセット中の正解ラベルを用いて、「どちらの方が難しいか」をラベルとして付与することで作り直した訓練データを用意する必要がある。そのため、この分類は教師あり学習の枠組みに入る。ランキング学習の性能評価については、正解データセットとの順位相関係数を用いて評価する事が可能である。また、文献 [32] では、均衡コーパス中のテキストの相対順位を用い

ることによって、テキストの難度の尺度そのものを表す研究が行われている。

回帰式 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) [18] や、SMOG grade[25]、Coleman-Liau index [3] などがこれにあたる。また、同様の回帰式を用いたアプローチは英語以外の言語でも広く行われている。日本語のリーダビリティについても、文献 [15] のグループによって継続的に研究が行われている。

こうした回帰式による手法は、通常、評価用データセットの正解ラベルを用いて訓練し、識別器を構成するといった手順を経ずに、回帰式のみが手法として示される。回帰式の性能評価は、評価用データセットを用いて行うには、評価用データセット中の各テキストに対して、回帰式から難度のスコアを計算し、これを正解ラベルと照らし合わせ、順位相関係数を用いて行うことができる。この際、難度のスコアは、回帰式であるため、通常、同順は少ないのに対し、正解ラベルでは、同じラベルが振られたテキストは全て同順であるため、同順補正が必要となり、どのような同順補正の方法を用いたのかを論文に明記するべきであるが、同順補正について論文中で言及していない研究もある。

これらの手法では、回帰式自体は、回帰問題を解くことによって求められているため、教師あり学習であるように見えるが、実際には、回帰問題を解く際に用いられたデータセットは、評価用のデータセットとは全く別のデータセットが使われていることが多い。自然言語処理分野においては、生テキストデータのみを用いて、アノテーションによる正解ラベルを用いないアプローチは「教師なし」と表現する事が多い。評価用データセットの正解ラベルを全く用いていないという意味では、回帰式を用いたアプローチは教師なし学習とみなすことができる。実際、後述の文献 [24] では、評価用データセットとは異なるコーパスを用いて学習された言語モデルからのスコアを「教師なし（unsupervised）」と表現している。

言語モデル 直近で発表された論文 [24] では、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し、この順でテキストを並び替える手法が「教師なしリーダビリティ判定」として紹介されている。あらかじめ対象言語のコーパスで訓練された言語モデルの情報は使用しているので、転移学習やドメイン適応の一種とみなすこともできる。いずれにせよ、データセット中の正解ラベル情報は一切使用しない設定であるため、教師なし学習の一種とみなすことができる。

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 [19], [27]。

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 [2], [19], [27]。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。このように、個人化リーダビリティは、学習者からの一次情報を重視した手法と言える。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 [2] 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 [26]。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 [10], [11]。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 [8] 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 [6]。

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 [22]。

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 [21] 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 [7]。

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 順位相関による自然言語処理分野のリーダビリティ尺度との比較

前述のように、個人化リーダビリティは、自然言語処理分野の語学教員がつけたラベルを予測するタイプのリーダビリティ異なる思想に基づいてはいるものの、大まかには、両者の傾向は一致する事が多いように思われる。自然言語処理分野のリーダビリティ手法と、個人化リーダビリティの手法を比較するにはどうしたらいいだろうか？

個人化リーダビリティは、所与のテキストに対して、各読み手にとってのリーダビリティを返す手法であるが、これを読みかえると、各読み手ごとに 1 つのリーダビリティの判定器を構成しているとも考えられる。ある 1 人の読み手に注目したとき、所与のテキストに対して、その既知語率や、既知語率が閾値を超える確率を、そのテキストのリーダビリティとしてみなしてしまう方法が考えられる。例えば、語彙テスト結果データセット中で最も標準的な語彙力の学習者にとっての個人化リーダビリティを、一般的なリーダビリティ尺度として用いることが考えられる。このようにすれば、個人化リーダビリティから、典型的なリーダビリティの問題設定のリーダビリティ判定器を作成することが可能である。

こうして個人化リーダビリティから作成した各読み手ごとのリーダビリティ判定器は、リーダビリティ評価用データセットの正解ラベルを一切使わず、語彙テストデータのみから構成できるため、表 1 の分類に従えば、「教師なし」の手法の 1 種とみなせる。また、既存の自然言語処理のリーダビリティ評価用データセットとの比較は、他の手法同様、順位相関を用いて行える。

3.5 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 [2] 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 [26]。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。

[6] 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 [1] 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。

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

(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、$\mathrm{sigmoid}(x) = \frac{1}{1+\exp(-x)}$ 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。$\mathrm{sigmoid}(0) = \frac{1}{1+1} = \frac{1}{2}$ 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。

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

(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、$\mathrm{freq}_k(v)$ 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 $\mathrm{freq}_k(v)$ 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 [*2]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 [*3]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い。ここで、$v \in \mathcal{T}$ は、テキスト中の単語 v を表す。

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

また、$p(z=1|v,l_{\mathrm{avg}})$ が計算できれば、(3) の代わりに、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 [7]。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 [29]。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 [*4]。このパッケージには、英語のリーダビリティ判定式とし

[*2] https://scikit-learn.org/stable/
[*3] https://www.csie.ntu.edu.tw/~cjlin/liblinear/
[*4] https://pypi.org/project/readability/

て、Flesch-Kincaid Grade Level[18], ARI, Coleman-Liau Index, Flesch Reading Ease, Gunning Fog Index, LIX, SMOG Index [25], 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 [24] 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。[24] 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[*5]の **sent_tokenize** 関数を用いた。

[24] 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、[24] 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、[24] 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 [24] 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 [24] 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 [6] 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 [*6]と Corpus of Contemporary American English (COCA) [*7]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。[24] 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 ρ、Kendall's τ を用いた。さらに、今回は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。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。しかし、5 件法と 10 件法を比較する場合など、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り[*8]、今回は、こちらの値も表示した。

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、提案手法が全ての尺度で最も良い性能を示した。提案手法は 0.730 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 に近い順位相関を達成した。

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、提案手法のスコアに 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(**提案手法のスコア**) 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 [24] よりも大きな事前学習モデルを用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 [24] 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。[24] 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

[*5] nltk.org
[*6] https://www.english-corpora.org/bnc/
[*7] https://www.english-corpora.org/coca/
[*8] https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient

表 2 OneStopEnglish データセットでの実験結果・考察

教師あり/なし	手法	Spearman's ρ	Kendall's τ-b	Kendall's τ-c	Pearson's ρ
教師なし	Flesch-Kincaid	0.324	0.253	0.308	0.359
	ARI	0.317	0.248	0.302	0.351
	Coleman-Liau	0.373	0.295	0.359	0.372
	FleschReadingEase	-0.387	-0.301	-0.366	-0.426
	GunningFogIndex	0.331	0.257	0.313	0.362
	LIX	0.348	0.273	0.332	0.383
	SMOGIndex	0.456	0.360	0.438	0.479
	RIX	0.437	0.340	0.414	0.462
	DaleChallIndex	0.495	0.387	0.472	0.506
	TCN RSRS-simple	-	-	-	0.615(*)
	BERTLMavg	-0.220	-0.173	-0.210	-0.040
	BNC	-0.012	-0.009	-0.010	-0.006
	COCA	0.018	0.016	0.020	0.039
	提案手法	**0.730**	**0.592**	**0.709**	**0.715**
	exp(**提案手法のスコア**)	**0.730**	**0.592**	**0.709**	0.260
教師あり	spvBERT_half	0.751	0.729	0.725	0.747
	spvBERT	0.866	0.856	0.854	0.864

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。**

5. まとめ

本研究では、応用言語学と自然言語処理のリーダビリティ判定に対するアプローチの違いを、何に信を置くか、の違いから解説した。さらに個人化リーダビリティ判定の手法を用いて、一般的な自動リーダビリティ判定を行う手法を示し、語彙テストデータを用いて、「第二言語学習者にとっての単語難易度」を複数のコーパスの単語頻度を組み合わせて正確に求めることで、文脈を考慮した大規模言語モデルよりも教師なし設定では高い精度を得ることができることを示した。

大規模言語モデルのパープレキシティは、計算負荷が大きく携帯端末などでは判定が事実上難しいが、提案手法は単に 2 特徴量を用いたロジスティック回帰であるので、計算負荷の観点からは携帯端末で実行可能であると考えられる。今後の展望として、実際にリーダビリティを軽量・教師なしで判定できるアプリケーションを作成する事が挙げられる。こうした詳細については、自動リーダビリティ判定についての詳細は、`http://yoehara.com/readability/` にまとめる予定である。

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] Baker, F. B.: *Item Response Theory : Parameter Estimation Techniques, Second Edition*, CRC Press (2004).
[2] Beglar, D. and Nation, P.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).
[3] 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).
[4] Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research, *ITL-International Journal of Applied Linguistics*, Vol. 165, No. 2, pp. 97–135 (2014).
[5] 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).

[6] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).

[7] 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).

[8] 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).

[9] 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).

[10] 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.

[11] 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).

[12] 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).

[13] Flesch, J.: Flesch-Kincaid readability formula (1965).

[14] 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.

[15] 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).

[16] 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).

[17] Imperial, J. M.: Knowledge-Rich BERT Embeddings for Readability Assessment, *arXiv:2106.07935 [cs]*, (online), available from ⟨http://arxiv.org/abs/2106.07935⟩ (2021). arXiv: 2106.07935.

[18] 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).

[19] Laufer, B.: What percentage of text-lexis is essential for comprehension, *Special language: From humans thinking to thinking machines*, Vol. 316323 (1989).

[20] 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 (2010).

[21] 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).

[22] 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).

[23] 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).

[24] 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 (online), available from ⟨https://doi.org/10.1162/coli$_{a}$0398⟩ (2021).

[25] Mc Laughlin, G. H.: SMOG grading-a new readability formula, *Journal of reading*, Vol. 12, No. 8, pp. 639–646 (1969).

[26] Nation, I.: How Large a Vocabulary is Needed For Reading and Listening?, *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82 (2006).

[27] Nation, P.: *Teaching and Learning Vocabulary*, Heinle and Heinle, Boston, MA (1990).

[28] 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).

[29] 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).

[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] 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).

[32] 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).