計算機科学を学ぶのに必要な英語はどの程度難しいのか？
Difficulty of Computer Science Texts

江原 遥
Yo EHARA
東京学芸大学　教育学部
Faculty of Education, Tokyo Gakugei University
Email: ehara@u-gakugei.ac.jp

あらまし： 計算機科学分野では論文からマニュアルまで英語が支配的な言語であるため，非英語圏で計算機科学を学ぶ者は英語も同時に学ばなければならない問題がある．しかし，計算機科学で用いられる英語の英語学習者にとっての難しさは先行研究に乏しい．本研究では，実際に英語教師が評価したデータセットに基づき，英語学習者にとっての英文の可読性を自動判定する自動判定器を構成することで，計算機科学で用いられる英語がどの程度英語学習者にとって難しいのかを評価した．
キーワード：計算機科学，技術文書，可読性

1　はじめに

計算機科学では，学術論文はもちろん，ソフトウェアを開発・利用するのに必要なマニュアル等技術文書も英語が一次情報である．非英語圏の計算機科学を学ぶ学習者は英語も同時に学ぶ必要があるが，第二言語として英語を学ぶ学習者（英語学習者）にとって，計算機科学のテキストが英語としてどれだけ難しいのかは先行研究に乏しい．本研究では，実際に英語教師が評価したデータセットに基づき，英語学習者にとっての英文の可読性を自動判定する自動判定器を構成して，こうした文書がどの程度英語学習者にとって難しいのかを評価した．本研究は CogSci 2022 poster full paper に採択された [8]．以後，計算機科学 (Computer Science) を CS と略す．

2　リーダビリティ自動評価器の構築

ここでは筆者による研究 [7] を参考にリーダビリティ自動評価器の構築方法を説明する．リーダビリティの自動評価器の構築には，リーダビリティを人手評価した既存のデータセット（OneStopEnglish[13]）を用いた．OneStopEnglish を選択した理由には，(英語母語話者の子供などではなく）英語学習者にとっての可読性が評価されていること，公開されており入手が容易なこと，平均文長などの非本質的な特徴量から可読性が予見できる問題等，それ以前の研究で知られた問題に対策されていることが挙げられる．

このデータセットでは，各テキストに初級（elementary），中級（intermediate），上級（advanced）の 3 段階の何れかの可読性ラベルが付与されている．各テキストの出自は，Guardian 紙である．このデータセットでは，新聞記事のジャンルからテキストの可読性がわかってしまわないよう，元のテキストを英語教師が人手で上級・中級・初級に書き直している．各段階に 189 件，計 567 件のテキストがある．これをランダムに 339 件の訓練，114 件の検証，114 件のテストデータに分割した．

BERT[5] は，大規模な母語話者コーパスで事前訓練（pretraining）することで言語の基本的な構造を認識できるニューラルな言語モデルを作成し，その後に今回の可読性の分類など，個々のタスクにあわせて微調整 (fine-tuning) と呼ばれる訓練を追加することで，従来の機械学習より高い性能が報告されている手法である．事前訓練モデル **bert-large-cased-whole-word-masking**（`https://huggingface.co/models`）を用い，10^{-5} の学習率の Adam 法により 10 エポックで訓練した BertForSequenceClassification を用いた．

また，**Vocabulary-based** は，[6, 9] の方法を参考に，英語教育で用いられる英語の均衡コーパスとして標準的な British National Corpus (BNC)[1] と Corpus of Contemporary American English (COCA)[4] を特徴量にして作成した，ロジスティック回帰に基づく（可読性ラベルを訓練に用いないという意味で）教師なし自動リーダビリティ評価器である．これは，主に日本語母語話者の英語学習者である（他のデータセットとは独立な）被験者に単語テストを解いてもらったデータ [6] を用いて，この単語テスト被験者中の平均的な学習者がテキスト中の単語を全て知っている確率をテキストのリーダビリティとした（より厳密には確率値の対数負値）．

表 1: リーダビリティ自動評価器の性能．**spvBERT** のみ可読性ラベルを用いる教師あり設定．

手法	スピアマンの順位相関係数	相関係数
Flesch-Kincaid	0.324	0.359
ARI	0.317	0.351
Coleman-Liau	0.373	0.372
FleschReadingEase	-0.387	-0.426
GunningFogIndex	0.331	0.362
LIX	0.348	0.383
SMOGIndex	0.456	0.479
RIX	0.437	0.462
DaleChallIndex	0.495	0.506
TCN RSRS-simple	-	0.615(*)
Vocabulary-based	**0.730**	**0.715**
spvBERT	0.866	0.864

既存の古典的なリーダビリティ評価式の実装には，**readability** ライブラリを用いた．これには，**Flesch-Kincaid** (Flesch-Kincaid Grade Level, FKGL) [11]，**ARI** (Automated Readability Index)，**Coleman-Liau** Index [2]，**Flesch Reading Ease** [10]，**Gunning Fog Index**，**LIX**，**SMOG Index**，RIX index，**Dale-Chall Index** [3] が実装されている．紙面の制限のため代表的な手法以外の参考文献や各手法の評価式の数式など詳細は，`https://pypi.org/project/readability/` を参照されたい．

表 1 に結果を示す．数値はスピアマンの順位相関係数，相関係数であり，高いほど人手評価との相関が高く，性能が良い．教師なし設定では，提案手法である **Vocabulary-based** が，既存の **TCN RSRS-simple**[12] より高い性能を示すこと，教師あり設定の **spvBERT** が教師なし設定の他の手法より高い性能を示すことがわかる．**spvBERT** も **Vocabulary-based** も英語教師による人手評価と高い相関を示すことから，英語学習者にとっての可読性の自動評価による調査に有用であるとわかる．

3　計算機科学のテキストによる実験

計算機科学関連のテキストとして，GitHub と ACL Anthology の 2 つのサイトからテキストを取得した．ACL Anthology は，自然言語処理の論文を概要も含めて多数掲載している．そこで，学術論文のソースとして ACL Anthology を選んだ．入手した全概要の中から無作為に選んだ 1,000 件の概要を使用した．また，もう一つの学術論文のソースとして，PubMed のウェブサイト（`https://pubmed.ncbi.nlm.nih.gov/download/`）から 55,410 の abstracts を入手し，その中からランダムに 1,000 件の概要を入手し，使用した．

対象とするソフトウェアマニュアルの収集にあたっては，市販のソフトウェアのマニュアルは分析対象から除外した．市販のソフトウェアのマニュアルは通常，専門

の校正会社によって校正されているため，一般的なテキストの傾向を表すというよりは，単に使用している校正会社の校正基準が明らかになるだけだからである．

一方，オープンソースソフトウェアの場合，英語学習者と英語母語話者の両方がソフトウェア開発に深く関わっており，オープンソースコミュニティで開発されたソフトウェアマニュアルの読みやすさの基準は通常存在しない．これは，オープンソースソフトウェアであっても，ソフトウェアマニュアルの構造には多くのルールが存在することと対照的である．そこで，本研究の分析対象として，オープンソースソフトウェアのホスティングサイトである GitHub を選択した．GitHub には多くのプロジェクトが登録されているが，何年もメンテナンスされていないソフトウェアや，一人の開発者によって開発されたソフトウェアリポジトリも多く存在する．このようなソフトウェアリポジトリは，英語学習者のためのソフトウェアマニュアルの読みやすさに関する本研究の対象外であることは明らかである．したがって，このようなソフトウェアリポジトリは分析から除外した．具体的には，2021 年 11 月から 2022 年 1 月にかけての，毎月の活動量トップ 10 件のリポジトリの README.md を対象とした．(`https://github.com/trending?since=monthly`)．Markdown 形式である README.md に対して，これを単純にテキストに変換した (Raw) と，テキスト変換の際に Markdown 中で使い方の例示など通常プログラミング言語の記述に使われるコードブロック部分を削除した (Code Removed) の 2 種類のテキストを用意した．

OneStopEnglish [13] データセットを用いて，前述の **spvBERT** の分類器を収集したコーパスに適用し，その出力をもって CS テキストのリーダビリティを評価した．表 2 に結果を示す．表中の数値は比率であり，各行を足すと 1 になる．全体的な傾向としては，GitHub は過半数のテキストが中級と判定されているのに対し，ACL Anthology は上級と判定されたテキストが多く，明らかに難しいことが分かる．なお，“初級”, “中級”, ‘‘ 上級” の定義は OneStopEnglish データセット [13] の定義に従う．ACL Anthology と GitHub (Code Removed)，ACL Anthology と GitHub (Raw) の間は，統計的に有意な差があった（Mann-Whitney U test, $p < 0.01$ ）．この結果から，ACL Anthology は GitHub (Raw) や GitHub (Code removed) よりも難易度が高いことが明らかになった．一方，GitHub (Code removed) と GitHub (Raw) の間には統計的な有意性は認められなかった．これは，コード削除の効果が限定的であることを示唆している．なお，比較のため，医療・生物系の論文からなる PubMed についても概要の可読性を **spvBERT** で計測し，その比率を示した．PubMed は ACL ANthology よりもさらに難しいと判定されていることが分かる．

spvBERT ではなく，学習者の語彙テスト結果データを用いてデータセット中の平均的な学習者がテキスト中のすべての語を知っている確率からスコアを計算する **Vocabulary-based** でも，同様に，GitHub は ACL Anthology より易しく評価された．GitHub の読みやすさの平均スコアは 0.117（Code removed），ACL Anthology は 0.140 であり，統計的有意な差がみられた（Mann-Whitney 検定，$p < 0.01$）.（スコアが高いほど，読みにくいことを意味する.）これは，英語学習者にとって計算機科学の論文の学術的な文章は特に難しいのに対し，ソフトウェアのマニュアルではそのような学術用語はほとんど使われないためと推測される．**Vocabulary-based** 評価器の利点は，**spvBERT** 評価器と異なり，実際にどのような語が英語学習者に難しいのかを出力することができる点である．例えば，GitHub のテキストでは *blockchain* や *automerge* といった語が，ACL Anthology では， *lexicosemantic* と *colingual* といった語が英語学習者に特に難しいと判定された．

表 2: 英語学習者にとっての可読性結果

-	初級	中級	上級
GitHub (Raw)	0.056	0.778	0.167
GitHub (Code Removed)	0.083	0.861	0.056
ACL Anthology	0.030	0.413	0.557
PubMed	0.005	0.189	0.806

4　まとめ

計算機科学の論文の概要の多くは，中級の英語学習者には読めないことを示した．一方，計算機科学のソフトウェアのマニュアルは英語学習者でもほとんど読めることを示した．このことは，英語学習者が計算機科学の論文を読解するには支援が必要であることを示す一方，[13] の定義による中級以上の英語学習者がソフトウェアマニュアル等を読解する上では，あまり支援が必要ない可能性を示唆している．本研究で収集した README.md，実験に用いたコード等は，`http://yoehara.com/` または `http://readability.jp/` で公開を検討している．

謝辞 本研究は，JST 戦略的創造研究推進事業（ACT-X, JPMJAX2006）の支援を受けた．

参考文献

(1) BNC Consortium. The british national corpus, version 3 (bnc xml edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium `http://www.natcorp.ox.ac.uk/`, 2007.
(2) Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, Vol. 60, No. 2, p. 283, 1975.
(3) Edgar Dale and Jeanne S Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pp. 37–54, 1948.
(4) Mark Davies. The corpus of contemporary american english (coca). Available online at `https://www.english-corpora.org/coca/`, 2008.
(5) 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.
(6) Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.
(7) Yo Ehara. Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In *Proc. of ICTAI*, pp. 806–814, 2021.
(8) Yo Ehara. Neural language model-based readability assessment of computer science introductory texts for english-as-a-second language learners. In *CogSci poster full paper*, 2022.
(9) Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *J. of Information Processing*, Vol. 26, pp. 267–275, 2018.
(10) Rudolf Flesch. A new readability yardstick. *J. of Applied Psychology*, Vol. 32, No. 3, pp. 221–233, 1948.
(11) J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Tech. Training Command Millington TN Research Branch, 1975.
(12) Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, Vol. 47, No. 1, pp. 141–179, April 2021.
(13) Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proc. of BEA*, 2018.