Readability of Scientific Papers for English Learners in Various Fields of Science

Yo Ehara
Tokyo Gakugei University
ehara@u-gakugei.ac.jp

Abstract

Most scientific papers are written in English. Non-native English speakers must simultaneously learn English and their area of scientific expertise. This is an obstacle for non-native English speakers in studying science. The difficulty of English used in scientific papers varies from field to field. For example, the vocabulary used in the medical sciences seems apparently difficult for English learners. Which scientific fields use English that is difficult for English learners to what extent? There are few existing studies in this regard. In this study, we compare the readability of English papers in various scientific fields by constructing and applying state-of-the-art artificial intelligence-based automatic readability assessor trained using actual textual data obtained by testing language learners and collecting judgments of language teachers. In experiments, our automatic assessors confirmed the intuition that medical science papers tend to be difficult for English learners. Moreover, our automatic assessors successfully quantified which field is difficult for English learners to what extent.

1 Introduction

English is one of the languages used by most scientific publications; it is also a second language for many scientists and those who learn science [12]. Hence, the readability of scientific publications for English as a second language (ESL) learners is essential for determining and developing the support that ESL learners need to learn science. If the language gap between native speakers and ESL learners causes misunderstandings in the interpretation of scientific papers, it will significantly hinder the development of science. However, few studies have investigated this issue as we discuss in the Related Work section.

To this end, in this paper, we assessed the readability of scientific publications for ESL learners. The readability (as an English text) of the main body of a paper is too technical to be properly evaluated, even for humans. Instead, we targeted the readability of the title and abstract, which are typically used to decide whether to read the main body of the paper.

To avoid biasing our analysis to one particular field, we obtained abstracts from the databases of two different fields: We analyzed 55,410 abstracts taken from PubMed for medical and life sciences, and 27,686 abstracts taken from the ACL Anthology for natural language processing. As a large-scale costly manual readability assessment is impractical, we constructed two contrastive automatic readability assessors: BERT-based assessor, and Vocabulary-based assessor with high interpretability.

Rather than simply predicting text readabilities, one of this paper's main contributions lies in the vocabulary-based assessor that can output the list of words difficult for ESL learners. By using the list, we can better understand which domain-specific terms and knowledge make scientific papers difficult for ESL learners.

NeurIPS 2022 AI for Science Workshop. This paper is non-archival. This workshop does not have official proceedings.

2 Related Work

Regarding studies that deals with improving the readability of scientific papers, there are two lines of previous studies. While both lines of studies do not deal with ESL learners and hence their goals are different from ours, we briefly explain them because they deal with scientific papers' readability in general. In one line of studies [17, 2], they try to summarize scientific papers textually or visually to read many papers quickly. In another line of studies [15, 1], they study methods to improve the typesetting and rendering of scientific papers such as in a Portable Document Format (PDF). Similar to this line of study, an eye-tracking movement study was conducted on scientific papers [16] to find the double-column format is superior over single-column format in terms of subjects' fluency and fastness in reading.

One study is closely related to ours [3], so we briefly explain their work and difference from ours. In their work, in order to "see how readability affects scientific impact" such as citations, they apply traditional methods for assessing readability such as the Flesh-Kincaid readability score [11] to the abstracts of scientific papers.

While we also apply readability assessments to the abstracts of scientific papers, there are some notable differences from the work [3]: First, the goal is different: they do not deal with readability for ESL learners in [3]. Second, the genres of papers' abstracts are limited to informatics-related terms such as "artificial intelligence" and "Smart contracts", not including scientific fields other than informatics. Third, they applied only traditional readability scoring methods to the abstracts of scientific papers and did not propose or use methods that suit for measuring the readability for ESL learners as we will do in this paper.

In terms of readability for second language learners, the relationship between the words that a second language learner knows and the readability of a text for the learner was previously investigated in [8].

3 Automatic Readability Assessment

This section formalizes the problem of automatic readability assessment. Let us suppose that we have N texts to assess: we write the set of texts as $\{\mathcal{T}_i | i \in \{1, \ldots, N\}\}$. Let $\mathcal{Y}$ be the set of readability labels. Labels are typically ordered in the order of difficulty. For example, in the *OneStopEnglish* dataset [18], we can set $\mathcal{Y} = \{0, 1, 2\}$, where 0 is elementary, 1 is intermediate, and 2 is advanced. The number of levels depends on the evaluation corpus. Using $\mathcal{Y}$, we write the label for $\mathcal{T}_i$ as $y_i \in \mathcal{Y}$.

3.1 Goal in Unsupervised Setting

Given each text $\mathcal{T}_i$, an *assessor* outputs its readability score s_i. In a supervised setting, the *assessor* knows the number of levels in the evaluation corpus from training examples. Hence, s_i ranges within $\mathcal{Y}$: $s_i \in \mathcal{Y}$. However, in an unsupervised setting, it is noteworthy that the assessor does not know $\mathcal{Y}$, or how many levels the evaluation corpus has, because no label is given. Hence, even if only integers are allowed for y_i, s_i can be a real value.

Throughout this paper, we write arrays using [and]. Given N texts $[\mathcal{T}_i | i \in \{1, \ldots, N\}]$, our goal is to make an assessor output arrays of readability scores $[s_i | i \in \{1, \ldots, N\}]$ that *correlate well* with the array of labels $[y_i | i \in \{1, \ldots, N\}]$. Here, there are multiple types of correlation coefficients between the array of scores and the array of labels, which we explain in the later sections. Typically, we should use *rank coefficients* such as Spearman's ρ, defined as the Pearson's ρ between rankings, when s_i is real-valued.

3.2 BERT-based supervised assessor

The former assessor was trained on a standard dataset of text readability [18] for ESL learners. Each text was annotated by a language teacher. This assessor is based on bidirectional encoder representations from transformers (BERT) [6], which is the current standard for building highly accurate text classifiers, and takes textual contexts into account for the assessment. Although this assessor improves accuracy, it has low interpretability of the assessment results, and it is difficult to interpret what types of words are particularly difficult for ESL learners.

3.3 Vocabulary-based Unsupervised Assessor

The latter assessor is based on [9]. The latter assessor is trained solely on a dataset of the vocabulary test results of ESL learners [7]; hence, although this assessor uses supervised learning from the viewpoint of machine-learning, it is categorized as an *unsupervised* readability assessor by [14] as it does not use any text readability labels. Notably, this assessor is solely based on vocabulary; hence, it cannot consider textual contexts. Although it does not achieve the highest accuracy, its results are highly interpretable because it can show which word in the text is difficult to what type of learner.

To perform the assessment, this assessor calculates the bag-of-words probability that the average ability test-taker knows all the words in a given text and regards its negative logarithm as the readability of the text. In other words, this assessor is personalized. As will be explained later, it uses a parameter that can be interpreted to represent the ability of each language-learner who takes the test; thus, it considers the knowledge level of each learner. Additionally, the weights of different words are obtained by considering word frequencies from multiple corpora.

We used questions from the vocabulary size test, a widely used vocabulary test in applied linguistics [5]. Each question asks about a word in a multiple-choice question format. The test consists of 100 questions. [7] used this test to have 100 second-language learners take the test and to collect their responses. Their data were published and made publicly available. We used their dataset to train our classifiers.

We want to analyze vocabulary test results to obtain word difficulty values encoding learners' language knowledge. To this end, we employed the idea of *item response theory* [4], a statistical model that can estimate learners' abilities and test questions' difficulties from the learners' responses to the questions.

Let $\mathcal{V}$ be the set of vocabulary, and let $\mathcal{L}$ be the set of learners. Let $z_{v,l} \in \{0, 1\}$ be the result of whether learner $l \in \mathcal{L}$ correctly answered the question for word $v \in \mathcal{V}$: $z_{l,v} = 1$ if l answered correctly for word v; otherwise, $z_{l,v} = 0$. Correct answers usually imply that l knows word v. Then, by using $\{z_{v,l}\}$ as the training data, we train the following model:

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

In Equation 1, a_l is the ability parameter of learner l, d_v is the difficulty of word w, and sigmoid denotes the logistic sigmoid function, i.e., $\text{sigmoid}(x) = \frac{1}{1+\exp(-x)}$. As $\text{sigmoid}(0) = \frac{1}{1+1} = \frac{1}{2}$, when a learner's ability a_l is larger than the word difficulty d_v, the probability that learner l knows word v can be written as follows: $p(z = 1|v, l) > \frac{1}{2}$ in Equation 1. To estimate learner ability and word difficulty, $z_{v,l}$ is given as z in Equation 1 in the training phase.

In Equation 1, d_v denotes the word difficulty estimated from the vocabulary tests. Here, in addition to the word difficulty for the words within the vocabulary test, we also want to obtain word difficulty values for all words that may appear in the target language. To this end, we calculate d_v from the word frequency in large balanced corpora as follows.

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

Finally, we use the following formula to obtain the readability of given $\mathcal{T}_i$ by calculating the probability that the average learner knows all the words that appear in $\mathcal{T}_i$. Here, l_{avg} denotes the average test-taker in the dataset.

$$s_i = score(\mathcal{T}_i) = -\log\left(\prod_{v \in \mathcal{T}_i} p(z = 1|v, l_{\text{avg}})\right) \tag{3}$$

4 Experiments for the Accuracy of the Readability Assessors

Our study is based on the accurate assessors of the readability assessors. Hence, in this section, we show the results of experiments to confirm the accuracy of the readability assessors mentioned in previous sections.

We used the OneStopEnglish dataset [18] for the source of readability for second language learners because it is one of the newest, publicly available, and reliable. The dataset has three levels:

Table 1: Predictive Performance of Readability. Only **spvBERT** is supervised: the others are unsupervised.

Method	Spearman's ρ	Pearson's ρ
Flesch-Kincaid	0.324	0.359
TCN RSRS-simple	-	0.615(*)
Vocabulary-based	**0.730**	**0.715**
spvBERT	0.866	0.864

Table 2: Readability Assessment Results of Scientific Texts for ESL learners.

-	Elementary	Intermediate	Advanced
ACL Anthology	0.037	0.860	0.103
PubMed	0.006	0.639	0.305

elementary, intermediate, and advanced. All three levels have 189 texts each, 567 texts in total. We randomly split these texts into a *training* set consisting of 339 texts, a *validation* set consisting of 114 texts, and a *test* set consisting of 114 texts. The training and validation sets were used to train solely supervised methods for comparison. Unsupervised methods did not use the training and validation sets; they used only the test set.

4.1 Compared Methods

As the BERT-based sequence classification has been reported to achieve excellent results [6], we applied the standard BERT-based sequence classification approach involving pretraining and fine-tuning. For the pretrained model, we used **bert-large-cased-whole-word-masking** in the Huggingface models (`https://huggingface.co/models`). Then, we fine-tuned the model using the 339 training texts. We named this fine-tuned model **spvBERT**, in which “spv” denotes being supervised. For fine-tuning, we used the Adam optimizer [13] with a setting of 10 epochs and a 0.00001 training rate. For the implementation of conventional readability formulae, we used the **readability** PyPI package (`https://pypi.org/project/readability/`). Table 1 shows the results. **TCN RSRS-simple** is the previous unsupervised state-of-the-art [14]. We can see that **Vocabulary-based** and **spvBERT** achieved the highest scores, implying that our assessors are highly accurate considering the current state-of-the-art.

Experiments with scientific texts Despite these differences, interestingly, our experimental results showed that the assessments of these two assessor types were similar overall. First, for both databases, the former assessor's assessment was that the majority of the abstracts were readable to intermediate English learners. Here, the definition of intermediate follows the definitions by [18]. The results are shown in Table 2.

Second, both assessors judged that the abstracts retrieved from the ACL Anthology were easier than those retrieved from PubMed. For the former assessor, this is obvious from the aforementioned classification results into the three levels. For the latter assessor, while the average readability score for the ACL Anthology was 18.45, that of PubMed was 31.25: a larger score indicates that the text is more difficult. Both assessors' results showing that the ACL Anthology abstracts are easier than the PubMed abstracts were statistically significant (the Mann-Whitney tests, $p < 0.01$). This is presumably because medical terminology, which is primarily of Greek origin and frequently used in PubMed, was particularly difficult for ESL learners. The qualitative results by the latter vocabulary-based assessor also confirmed this tendency. For example, the following words were assessed to be particularly difficult for ESL learners in PubMed: *hemihydrate*, *engraftment*. In contrast, those in the ACL Anthology: *lexicosemantic*, *colingual*.

Table 3 shows the analysis of **Vocabulary-based** method: it shows the words classified as the most difficult for the average-skilled ESL learner in the dataset of [7]. More details of the classifier were explained in Section 3.3. For each of the 1,000 randomly chosen abstracts, Table 3 shows the most difficult word in each abstract by **Vocabulary-based**, along with the number of times the word was classified to be the most difficult in each abstract. For example, “embeddings” in the “ACL” column

Table 3: Ranking of words that appear as most difficult in $1,000$ abstracts.

Rank	ACL		PubMed	
-	Word	Frequency	Word	Frequency
1	embeddings	70	acyl	16
2	nlp	45	wa	11
3	datasets	26	acylated	7
4	semeval	21	dhap	6
5	dataset	14	arhl	6
6	pretrained	12	ipv	5
7	convolutional	11	acyltransferase	4
8	arabert	11	alloimmune	4
9	subtask	10	deprotonation	4
10	hahackathon	10	autophagy	4

indicates that it appears as the most difficult word in the 70 of the 1,000 ACL Anthology abstracts. The fact that completely different words appear as the most difficult words in ACL and PubMed clearly shows that the words difficult for ESL learners are *domain-specific*: most of the words in Table 3, such as "datasets" and "acylated", express scientific *conceptual knowledge* specific to each scientific domain. Also, Table 3 shows that the distribution is more "long-tail" in PubMed than in ACL: considering the shape of Zipf's law, this suggests that more variety of words appear as difficult words in PubMed than in ACL.

5 Conclusions

We showed that 10%-30% of the scientific texts are not readable to intermediate ESL learners, implying that they need assistance in reading scientific texts. In particular, one of the characteristic contributions of this paper is that we showed the results of a highly explanatory method (Vocabulary-based) that can detect the words that are particularly difficult for ESL learners.

Since this paper is a preliminary study, there is much future work. For example, we can look into the readability of scientific abstracts in more fields. In addition, it will be necessary to manually evaluate the readability of scientific abstracts by creating a dataset that actually measures the readability of scientific abstracts using crowdsourcing.

Acknowledgments and Disclosure of Funding

This paper is non-archival. This paper is based on [10]. This work was supported by JST ACT-X Grant Number JPMJAX2006, Japan. We used ABCI of AIST and miniRaiden of RIKEN for the computational resources. We appreciate the anonymous reviewers for their valuable and insightful comments.

References

[1] Rakefet Ackerman and Tirza Lauterman. Taking reading comprehension exams on screen or on paper? a metacognitive analysis of learning texts under time pressure. *Computers in human behavior*, 28(5):1816–1828, 2012.

[2] Lorenzo Amabili and Nicole Sultanum. Paper maps: Improving the readability of scientific papers via concept maps. *OSF Preprints*, 2021.

[3] Lennart Ante. Readability affects scientific impact: Evidence from emerging technology discourses. *BRL Working Paper Series*, (21), 2021.

[4] Frank B. Baker. *Item Response Theory : Parameter Estimation Techniques, Second Edition*. CRC Press, July 2004.

[5] David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, 31(7):9–13, 2007.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pages 4171–4186, Minneapolis, Minnesota, June 2019.

[7] Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.

[8] Yo Ehara. Uncertainty-aware personalized readability assessments for second language learners. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1909–1916. IEEE, 2019.

[9] Yo Ehara. Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 806–814. IEEE, 2021.

[10] Yo Ehara. Analyzing readability of scientific abstracts for esl learners. In *Companion Proc. of the 12th International Learning Analytics and Knowledge Conference (LAK2022, poster)*, pages 86–88, 2022.

[11] JR Flesch. Flesch-kincaid readability formula, 1965.

[12] Dennis Fung. Teaching science through home and second languages as the medium of instruction: a comparative analysis of junior secondary science classrooms in hong kong. *International Journal of Science and Mathematics Education*, 19:1609—-1634, 2021.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[14] Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179, 2021.

[15] Khaled Moustafa. Improving pdf readability of scientific papers on computer screens. *Behaviour & Information Technology*, 35(4):319–323, 2016.

[16] Seyyed Saleh Mozaffari Chanijani, Mohammad Al-Naser, Syed Saqib Bukhari, Damian Borth, Shanley E.M. Alleny, and Andreas Denge. An eye movement study on scientific papers using wearable eye tracking technology. In *2016 Ninth International Conference on Mobile Computing and Ubiquitous Networking (ICMU)*, pages 1–6, 2016.

[17] João Ulisses, Rebeca P Díaz Redondo, and Ana Fernández Vilas. Automatic aid system to enhance the readability of scientific papers. In *EDULEARN19 Proceedings 11th International Conference on Education and New Learning Technologies: Palma, Spain. 1-3 July, 2019*, pages 2416–2424. IATED Academy, 2019.

[18] Sowmya Vajjala and Ivana Lučić. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proc. of BEA*, pages 297–304, 2018.