To What Extent Does Lexical Normalization Help English-as-a-Second Language Learners to Read Noisy English Texts?

Yo Ehara
Tokyo Gakugei University
4-1-1 Nukuikita-machi, Koganei-shi, Tokyo 184-8501 Japan
ehara@u-gakugei.ac.jp

Abstract

How difficult is it for English-as-a-second language (ESL) learners to read noisy English texts? Do ESL learners need lexical normalization to read noisy English texts? These questions may also affect community formation on social networking sites where differences can be attributed to ESL learners and native English speakers. However, few studies have addressed these questions. To this end, we built highly accurate readability assessors to evaluate the readability of texts for ESL learners. We then applied these assessors to noisy English texts to further assess the readability of the texts. The experimental results showed that although intermediate-level ESL learners can read most noisy English texts in the first place, lexical normalization significantly improves the readability of noisy English texts for ESL learners.

1 Introduction

Noisy English texts are not only problematic for processing text but also difficult for humans to read, even for native speakers when, for instance, they lack the background to decipher abbreviations. How difficult are noisy English texts for English as a second language (ESL) learners?

If noisy English texts are too difficult for ESL learners to read, it may affect their behavior on various media such as social networking services (SNS). For example, it is easy to imagine that ESL learners do not follow English accounts. The lexical normalization tasks are important as they could potentially change the behavior of ESL learners by improving readability. In this sense, assessing the readability of noisy English texts for ESL learners is closely related to assessing the extent to which the language gap leads to social division.

To this end, we first sought to understand the readability of noisy English texts by building highly accurate readability assessors. We used two approaches to build the assessors. The first was based on the field of educational Natural Language Processing (NLP) (Vajjala and Lučić, 2018). Using a corpus that is standard in this field, we built a highly accurate readability assessor using deep learning methods, such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019).

The second approach was to conduct readability assessments based on the vocabulary of English learners. These methods have been well studied in applied linguistics, whereby considerable research has revealed that English learners need to know more than 95% of the words in a text to read and understand them (Nation, 2006; Laufer and Ravenhorst-Kalovski, 2010). The idea of assessing text readability via each learner's vocabulary knowledge is beneficial for interpreting the readability assessment results. Therefore, we also constructed a classifier based on a dataset of English learners' vocabulary tests (Ehara, 2018), to determine the number of words in a text that an English learner knows.

In experiments carried out on a standard dataset for evaluating readability (Vajjala and Lučić, 2018) in educational NLP, the results provided by the two approaches were in close agreement. The experiments using an English noisy text corpus also showed that most noisy English texts were readable by intermediate English learners. Using gold-normalized texts in the noisy English text dataset, we found that lexical normalization improves the text readability of second language learners. This improvement in readability was statistically significant in the vocabulary-based assessor's results.

The contributions of this study are as follows:

1. We evaluated the readability of noisy English texts for ESL learners by using accurate readability assessors.

2. We showed that intermediate ESL learners can read most noisy English texts.

3. We show that lexical normalization improves the readability of texts for ESL learners in both approaches, namely, the BERT readability assessor and the assessor based on the vocabulary of English learners. For the latter assessor, the result was also statistically significant.

2 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 $\{\mathcal{T}_i | i \in \{1, \dots, N\}\}$. Let $\mathcal{Y}$ be the set of readability labels. Labels are typically ordered in the order of difficulty. For example, in the *OneStopEnglish* dataset (Vajjala and Lučić, 2018), 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 $y_i \in \mathcal{Y}$.

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 $[\mathcal{T}_i | i \in \{1, \dots, N\}]$, our goal is to make an assessor output arrays of readability scores $[s_i | i \in \{1, \dots, N\}]$ that *correlate well* with the array of labels $[y_i | i \in \{1, \dots, N\}]$. 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 Vocabulary Testing-based Readability

Fig. 1 shows example questions from the vocabulary size test, a widely used vocabulary test in applied linguistics (Beglar and Nation, 2007). Each question asks about a word in a multiple-choice question format. The test consists of 100 questions like those shown in Fig. 1. Ehara (2018) 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.

```
15. deficit:
The company <had a large deficit>.
a: spent a lot more money than it earned
b: went down a lot in value
c: had a plan for its spending
                 that used a lot of money
d: had a lot of money stored in the bank
```

Figure 1: Examples of the Vocabulary Size Test (Beglar and Nation, 2007), one of the most widely accepted vocabulary tests to quickly assess language learners. They are asked to choose the option that paraphrases the part between "<" and ">" from a, b, c, and d.

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* (Baker, 2004), 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:

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

In Eq. 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., $\text{sigmoid}(x) = \frac{1}{1+\exp(-x)}$.

The logistic sigmoid function is the binary version of the softmax function, which is frequently used in neural classifiers. It is a monotonously increasing function ranging within $(0, 1)$. 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 Eq. 1. Likewise, by using Eq. 1, we can compare a learner's ability and word difficulty in the same dimension.

To estimate learner ability and word difficulty, $z_{v,l}$ is given as z in Eq. 1 in the training phase. In this way, in *item response theory*, learner ability and word difficulty are comparable, and these parameters are estimated from the test result data.

In Eq. 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:

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

In Eq. 2, K is the number of corpora to use, $\mathrm{freq}_k(v)$ denotes the frequency of word v in the k-th corpus, and w_k is the weight parameter of the k-th corpus. In summary, given the vocabulary test results $\{z_{v,l}\}$ and corpus frequency features $\mathrm{freq}_k(v)$, we can estimate the parameters: namely, the weight of the k-th corpus w_k and learner l's ability a_l. To implement the model, we used logistic regression, by following (Ehara, 2018). Note that this model does not use the valuable readability label $\{y_i\}$ in the training phase, so is unsupervised.

As shown in Eq. 1, we employed IRT-based modeling in this study. IRT-based modeling has been used in many previous NLP studies such as (Ehara et al., 2012, 2016; Ehara, 2019; Settles et al., 2020; Ehara, 2020).

After estimating the parameters using the above-mentioned procedure, we use the following formula to obtain the readability of given $\mathcal{T}_i$. Here, l_{avg} denotes the test-taker whose estimated ability parameter is closest to the average of the estimated ability parameter values $\{a_l\}$s. Intuitively, the following equation calculates the probability that the average learner knows all the words that appear in $\mathcal{T}_i$ and uses it as the readability score.

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

4 OneStopEnglish Experiments

We used the OneStopEnglish dataset (Vajjala and Lučić, 2018) for the source of readability for second language learners because it is one of the newest, publicly available, and reliable in the sense that no known trivial features are effective for predicting its labels such as average sentence length.

The dataset has three levels: elementary, intermediate, and advanced. The original articles were taken from the Guardian newspaper. The OneStopEnglish dataset is a parallel corpus, i.e, language teachers manually rewrote the original articles into the three aforementioned readability levels: hence, its readability labels are not easily estimated from the topics of texts.

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 (Devlin et al., 2019), 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 [1].

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 (Kingma and Ba, 2015) 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 [2]. We used almost all readability formulae implemented in this package for our experiments: namely, **Flesch-Kincaid** (Flesch-Kincaid Grade Level, FKGL) (Kincaid et al., 1975), **ARI** (Automated Readability Index) (Senter and Smith, 1967), the **Coleman-Liau** Index (Coleman and Liau, 1975), **Flesch Reading Ease** (Flesch, 1948), the **Gunning Fog Index** (Gunning, 1952), **LIX** (Björnsson, 1968), the **SMOG Index** (Mc Laughlin, 1969), the RIX index (Anderson, 1983), and the **Dale-Chall Index** (Dale and Chall, 1948). More details of these formulae and their implementation are described on the project page. All of these readability formulae are *unsupervised* in the sense that they do not require any training data.

The **Vocabulary-based** model was trained on a publicly available vocabulary dataset (Ehara, 2018). For the corpus word frequency, we used the frequencies taken from the British National Corpus (BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA) (Davies,

[1] https://huggingface.co/models

[2] https://pypi.org/project/readability/

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
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

2008). Both corpora are used extensively in applied linguistics in English as a rough measure for determining word difficulty (Nation, 2006).

4.2 OneStopEnglish Results

Tab. 1 shows the experimental results. In all unsupervised methods, **Vocabulary-based** achieved the best results in all rank correlation coefficients. **TCN RSRS-simple** is the best model on the OneStopEnglish dataset in Martinc et al. (2021). As they show only the performance measured by the Pearson correlation, we wrote – for Spearman's ρ. While a direct comparison is not possible as denoted by $(*)$, **Vocabulary-based** outperforms it.

Importantly, we can observe that both **Vocabulary-based** and **spvBERT** achieve high predictive performance. This result indicates that the two approaches to assessing readability derived results that were in close agreement.

5 Experiments with Noisy English Texts

For the experiment, we used the English training data of the W-NUT multilingual lexical normalization shared task [3]. We chose this dataset so that future studies on lexical normalization using a language other than English would have a baseline to compare against. The dataset consists of $2,361$ short noisy texts along with the corresponding lexically normalized texts. We selected $1,325$ pairs from this dataset, after removing the pairs that were identical before and after lexical normalization. To assess the effect of lexical normalization, we simply used the gold-lexically normalized texts of the dataset.

-	Elem.	Int.	Adv.
Before	0.486	0.512	0.002
After	0.495	0.503	0.002

Table 2: Readability Assessment Results for ESL learners before/after Lexical Normalization

-	-	After		
-	-	Elem.	Int.	Adv.
Before	Elem.	1051	97	0
	Int.	117	1091	0
	Adv.	1	0	4

Table 3: Matrix of Readability Assessment for ESL learners before/after Lexical Normalization

Tab. 2 shows the readability assessment results using **spvBERT** before/after lexical normalization. As most texts are elementary or intermediate, this result implies that *intermediate-level ESL learners can read most noisy English texts*. We can also see that the difficulty of elementary texts can increase based on lexical normalization. Tab. 3 displays the confusion matrix before/after lexical normalization by **spvBERT**. We observe that lexical normalization between elementary and intermediate levels can work both ways: while 117 texts were converted from elementary to intermediate, 97 texts were converted from intermediate to elementary. No text was converted to the advanced level. While Tab. 2 and Tab. 3 show that lexical normalization improves readability, the improvement was not statistically significant because 97 texts were reversely converted.

Next, we explain the readability assessment results of **Vocabulary-based**, which uses Eq. 3. The smaller the value, the easier it is for ESL learners to read. While the average score before lexical normalization was 4.00, the value after lexical normalization was 3.53. Fig. 2 is the scatter plot of the readability scores before/after lexical normalization. Fig. 2 clearly shows that the texts become significantly easier to read after lexical normalization (Wilcoxon test, $p < 0.01$).

[3] http://noisy-text.github.io/2021/multi-lexnorm.html

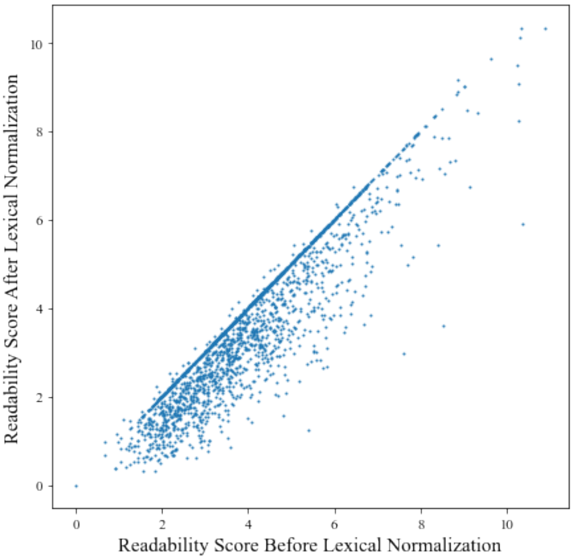

Figure 2: Eq. 3 Scores before/after Normalization

6 Discussion

In this study, we assessed the readability of text for ESL learners utilizing two sources of information. **spvBERT** is based on the annotations of the language teachers, and **vocabulary-based** is based on the results of the vocabulary test results of the learners. In this paper, both information from language teachers as well as learners is incorporated into the readability assessments described in the previous section.

However, there are limitations to automated readability assessment, and ideally a large-scale survey of ESL learners is required, which would be very costly and hence will be the subject of future work. Even in manual assessment, there are points to consider. For example, in annotating the readability of texts by language teachers, there is the issue of how well the language teachers grasp the characteristics of the students. In addition, even if learners themselves participate in assessing text readability, the cultural background of the learner, such as the influence of their first language (L1), may influence the performance.

7 Conclusions

We assess the readability of noisy English texts for ESL learners. We built highly accurate assessors using the two approaches. While intermediate ESL learners can read noisy English texts, lexical normalization improves readability for ESL learners, as was the case with our dataset.

Future work will include more comprehensive experiments using other datasets and evaluations by actual ESL learners.

Acknowledgment

This study was supported by JST ACT-X Grant Number JPMJAX2006 and JSPS KAKENHI Grant Number 18K18118. We used the ABCI infrastructure from AIST for the computational resources. We appreciate anonymous reviewers for their valuable comments.

References

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

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

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

C. H. Björnsson. 1968. Läsbarhet, Stockholm.

BNC Consortium. 2007. 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/.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Mark Davies. 2008. The corpus of contemporary american english (coca). Available online at https://www.english-corpora.org/coca/.

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

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

Yo Ehara. 2019. Neural rasch model: How do word embeddings adjust word difficulty? In *Computational Linguistics - 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11-13, 2019, Revised Selected Papers*, volume 1215 of *Communications in Computer and Information Science*, pages 88–96. Springer.

Yo Ehara. 2020. Interpreting neural CWI classifiers' weights as vocabulary size. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 171–176, Seattle, WA, USA → Online. Association for Computational Linguistics.

Yo Ehara, Yukino Baba, Masao Utiyama, and Eiichiro Sumita. 2016. Assessing Translation Ability through Vocabulary Ability Assessment. In *Proc. of IJCAI*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining Words in the Minds of Second Language Learners: Learner-Specific Word Difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. 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.

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

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, 22(1):15–30.

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

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

I. Nation. 2006. How Large a Vocabulary is Needed For Reading and Listening? *Canadian Modern Language Review*, 63(1):59–82.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.