LURAT: a Lightweight Unsupervised Automatic Readability Assessment Toolkit for Second Language Learners

Faculty of Education
Tokyo Gakugei University
Junior Assoc. Prof.
Yo Ehara

readability.jp yoehara.com

This research was supported by JST ACT-X Grant（JPMJAX2006）

Takeaways

- Automatic Readability Assessment (ARA)
  - The task of automatically assessing the readability of given texts for second language learners. The task can be regarded as a text classification task, however…
  - Problem: huge annotation cost of building training data. We need to hire language teachers and have them read and annotate many texts with readability labels.
  - To avoid such annotation cost, we want to build unsupervised assessors, which do not use the readability labels as training data.
  - Previous [Martinc et al., CL2021]: neural language model (LM)-based. LM-based fluency measures such as perplexities were used. No information taken from second language (L2) learners was used.
  - Ours: leverages information taken from L2 learners. We use vocabulary-test results of L2 learners to calculate word difficulty for L2 learners. Readability is modeled as the probability that an average L2 learner knows all words in the given text.
  - Experiments: ours outperforms previous in predictive performance, memory used for classification, and speed.

Evaluation Datasets for ARA

Old but famous readability formulae such as Flesch-Kincaid (1948):

they are unsupervised but built from evaluation datasets for ARA targeting children whose native language is English.

Evaluation datasets for ARA targeting English-as-a-Second Language (ESL) learners have been proposed recently:

- WeeBit [Vajjala and Meurers, 2012]
  - Weakness : classifiable not by readability but by text domains
- Newsela [Xu, Callison-Burch, and Napoles 2015]
  - Weakness: classifiable by merely using average sentence length
- OneStopEnglish [Vajjala and Lucic, 2018]
  - Do not have these weaknesses. Language teachers classify texts into elementary, intermediate, and advanced.

Vocabulary Test Result Dataset

```
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

26. malign:
His <malign> influence is still felt.
a: good
b: evil
c: very important
d: secret
```

Vocabulary Size Test [Beglar&Nation, 2007]

Multiple-choice test. Each question asks about a word. Consists of 100 questions.

Applied linguists use this test to vaguely assess the L2 vocabulary size of an L2 learner.

[Ehara, LREC2018] I previously built voc. test result dataset

100 learners x 100 words.

Data publicly available at:

yoehara.com

Proposed Method

Voc. Test Result Data [Ehara, LREC2018] 100 learners x 100 words

From the data, Proposed method is trained to estimate the prob. that each learner knows each word. Features: word frequencies

1. British National Corpus (BNC)
2. Corpus of Contemporary American English (COCA)

Probability that Learner *l* knows Word *v*:

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

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

z: 1 iff *l* knows *v*; 0 otherwise
a_l: ability parameter of *l*
d_v: difficulty parameter of *v*
$\text{freq}_k(v)$: freq. of word v in the *k*-th corpus
w_k: weight of word frequencies in *the k*-th corpus

Readability: prob. that a learner knows all words in the given text

Procedure: we first train the model and obtain

a_ls for all learners in the dataset

l_{avg}: the learner with a_l closest to the average of a_ls.

Readability is defined as follows. *T* is the given text.

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

This equation means that the readability score is defined as the (-log of the) probability that l_{avg} knows all words in Text *T.*

Experimental Setting

OneStopEnglish Dataset [Vajjala and lucic,2018]: language teachers manually classified texts (mostly news paper articles) into three levels: Elementary, Intermediate, Advanced

We have 567 texts in total. We split them into training, validation, and test set which consist of 339, 114, and 114 texts, respectively

Compared methods :

Conventional readability formulae (unsupervised) such as Flesch-kincaid

BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019]

pre-training: bert-large-cased-whole-word-masking

BERTLMavg: **no** fine-tuning, unsupervised. We use perplexity of BERT LM as readability score.

spvBERT: fine-tuned by OneStopEnglish dataset's training data split.

We used BertForSequenceClassification func. with Adam optimizer.

Evaluation Measures

Gold labels: discrete. Namely, elementary (0), intermediate (1), and advanced (2)

Example : [2, 1, 0, 0, 1]

Unsupervised readability scoring : continuous.

Example : [45.3, 39.2, 10.7, 13.2, 24.4]

How do we compare?

As the scores are not necessarily linearly increasing, the use of Pearson's ρ underestimates the predictive performance.

We should use rank-correlation measures with adjustments for ties (in ranking).

Namely, we should use Spearman's ρ or Kendall's τ-c

Kendall's τ-c is not implemented in scipy old versions.

Experimental Results 1/3

Supervision	Method	Spearman's ρ	Kendall's τ-b	Kendall's τ-c	Pearson's ρ
Unsupervised	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
	Proposed	**0.730**	**0.592**	**0.709**	**0.715**
	exp(**Proposed**)	**0.730**	**0.592**	**0.709**	0.260
Supervised	spvBERT_half	0.751	0.729	0.725	0.747
	spvBERT	0.866	0.856	0.854	0.864

Experimental Results 2/3

Supervision	Method	Spearman's ρ	Kendall's τ-b	Kendall's τ-c	Pearson's ρ
Unsupervised	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
	Proposed	**0.730**	**0.592**	**0.709**	**0.715**
	exp(**Proposed**)	**0.730**	**0.592**	**0.709**	0.260
Supervised	spvBERT_half	0.751	0.729	0.725	0.747
	spvBERT	0.866	0.856	0.854	0.864

Achieved better results from previous state of the art (not directly comparable as we could not obtain previous test set data)

Experimental Results 3/3

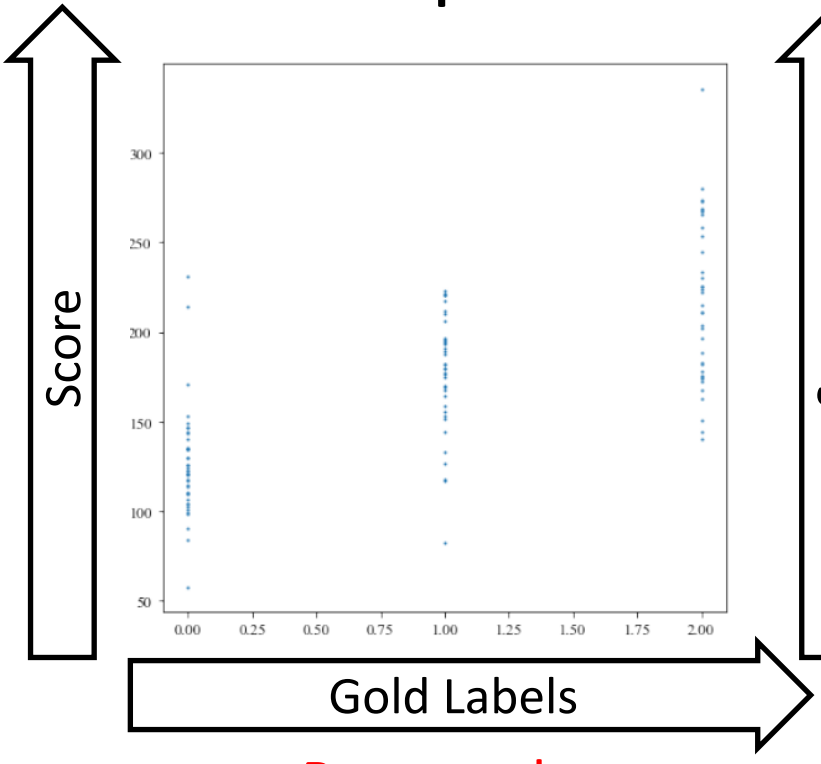
Supervision	Method	Spearman's ρ	Kendall's τ-b	Kendall's τ-c	Pearson's ρ
Unsupervised	Flesch-Kincaid	0.324	0.253	0.308	0.359
	ARI	0.317	0.248	0.302	0.351
	Coleman-Liau	0.373	0.295	0.359	0.372
	FleschReadingEase	-0.387	-0.301	-0.366	-0.426
	GunningFogIndex	0.331	0.257	0.313	0.362
	LIX	0.348	0.273	0.332	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
	Proposed	**0.730**	**0.592**	**0.709**	**0.715**
	exp(**Proposed**)	**0.730**	**0.592**	**0.709**	0.260
Supervised	spvBERT_half	0.751	0.729	0.725	0.747
	spvBERT	0.866	0.856	0.854	0.864

Raw word freq.s do not correlate well with manual readability labels

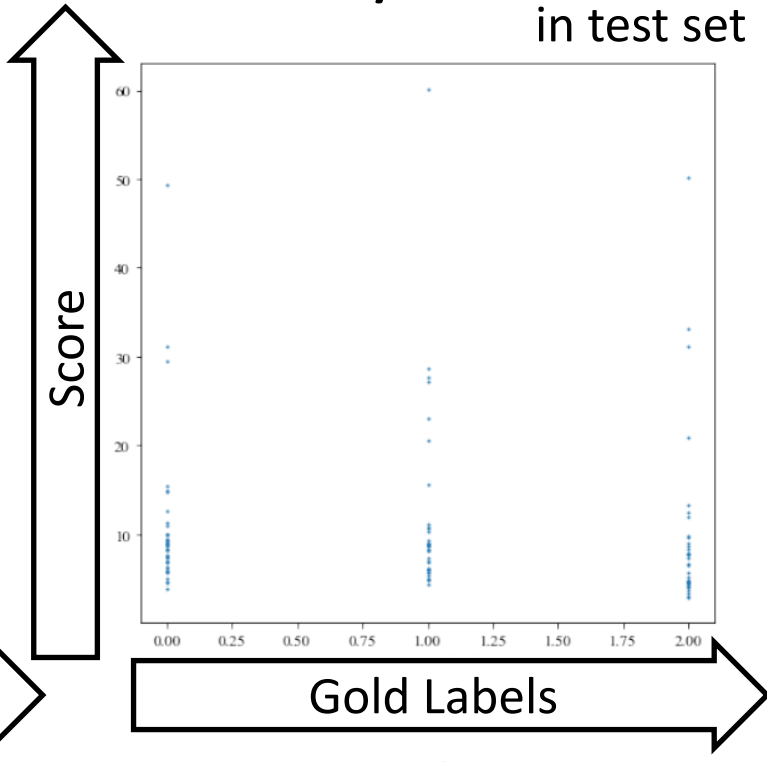
We need to estimate word difficulty

Detailed Analysis of Unsupervised Methods 1/2

Each dot: each text in test set



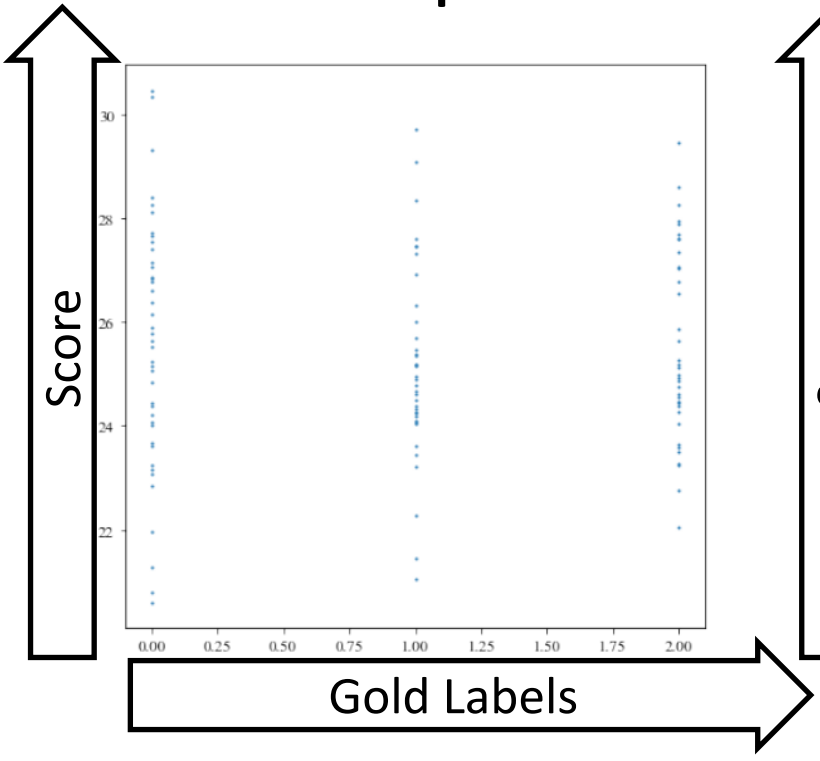
Proposed



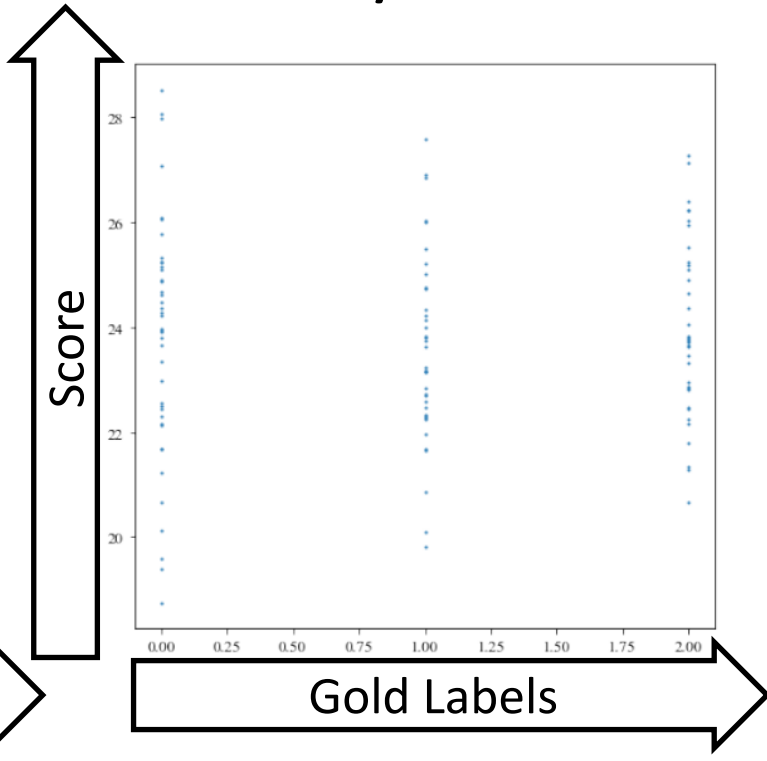
BERTLMavg

We can see that BERTLMavg more weakly correlate with gold labels.

Detailed Analysis of Unsupervised Methods 2/2



BNC



COCA

We can see that word frequencies do not correlate with gold labels.

Confusion matrices of supervised models

spvBERT

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

spvBERT_half

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

We can see that elementary/intermediate is easier to discern than intermediate/advance .

Memory and Speed Analysis

(Recap.) BERTLMavg: 0.220, Proposed: 0.730 (higher is better)

Memory:

BERTLMavg: requires 1,793 MiB GPU memory

Proposed: requires 0 GPU memory

10 MiB CPU memory

Classification time of unsupervised methods:

BERTLMavg: 368 seconds

Proposed: 5.37 seconds

Proposed outperforms BERTLMavg and works by far faster with by far lower memory.

Conclusions

- Unsupervised automatic readability assessment tasks for L2 learners.
- Previously no information taken from L2 learners was used. Neural LMs trained from texts written by native speakers were used.
- To make use of information taken from L2 learners efficiently, we proposed to use learners' vocabulary test results for readability assessment.
- Readability: the prob. that an average L2 learner knows all words in a text.
- Experiment results: our method outperformed large neural LMs in predictive performance.
- **Information taken from L2 learners was shown to be significant in ARA.**
- Our method is logistic regression and hence lightweight; it uses less memory and works on machines with low computational resources.

- Future work: JavaScript implementation of ARA tools that work on smartphones and websites.

Thank you for listening

readability.jp

yoehara.com