Unsupervised Sentence Readability Estimation Based on Parallel Corpora for Text Simplification

Rina Miyata[*]
Ehime University
miyata@ai.cs.ehime-u.ac.jp

Toru Urakawa
The Asahi Shimbun Company
urakawa-t@asahi.com

Hideaki Tamori
The Asahi Shimbun Company
tamori-h@asahi.com

Tomoyuki Kajiwara
Ehime University / The University of Osaka
kajiwara@cs.ehime-u.ac.jp

Abstract

We train a relative sentence readability estimator from a corpus without absolute sentence readability. Since sentence readability depends on the reader's knowledge, objective and absolute readability assessments require costly annotation by experts. Therefore, few corpora have absolute sentence readability, while parallel corpora for text simplification with relative sentence readability between two sentences are available for many languages. With multilingual applications in mind, we propose a method to estimate relative sentence readability based on parallel corpora for text simplification. Experimental results on ranking a set of English sentences by readability show that our method outperforms existing unsupervised methods and is comparable to supervised methods based on absolute sentence readability.

1 Introduction

Readability estimation of text, such as words, sentences, and documents, is applied to assist in text recommendation and simplification for a wide range of readers, including children (Xu et al., 2015), language learners (Xia et al., 2016), and people with cognitive disabilities (Yaneva et al., 2017), according to their language abilities. We work on readability estimation for sentences, which are the main units in the text simplification task (Alva-Manchego et al., 2020).

Since sentence readability depends on the reader's knowledge, objective and absolute readability assessments require costly annotation by experts. Therefore, corpora annotated with absolute readability are limited to a scale of $1k$ to $10k$ sentences even in English (Stajner et al., 2017; Arase et al., 2022), and are rarely available in other languages. This low-resource problem hinders research and development of high-quality supervised sentence readability estimation.

In this study, we train a relative sentence readability estimator based on labeled corpora for relative sentence readability, which are more accessible than those with absolute sentence readability labels. Our proposed method estimates which of two given sentences is more readable based on pairs of complex sentences and simpler sentences in parallel corpora for text simplification. The estimator is then applied to pairwise comparisons of a given set of sentences to rank them in terms of readability.

Experimental results on ranking a set of English sentences by readability show that the proposed method outperforms existing unsupervised methods. In addition, our proposed method achieved performance comparable to supervised methods that consider absolute sentence readability.

2 Related Work

For estimating text readability, supervised methods (Vajjala and Lučić, 2018; Deutsch et al., 2020) have been proposed that consider readability indices, linguistic features, and language model scores. Since they are based on corpora annotated with absolute sentence readability, they can not apply to languages without labeled corpora available.

Unsupervised methods such as FKGL (Kincaid et al., 1975) and other readability metrics and ranking methods based on relative readability estimation (Tanaka-Ishii et al., 2010) have been proposed. However, they are targeted at documents and are not applicable to sentences.

[*] Work done during an internship at The Asahi Shimbun Company.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 499–504
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

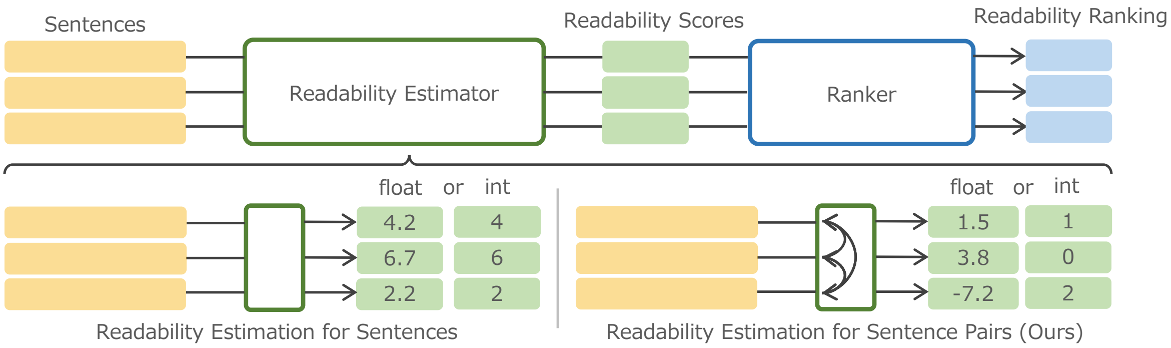

Figure 1: System overview

3 Method

We first train a relative sentence readability estimator that estimates which of two given sentences is more readable, based on parallel corpora for text simplification. The estimator is then applied to pairwise comparisons of a given set of sentences to rank them in terms of readability. The overall system process consists of estimating readability by using the Readability Estimator and then ranking them, as shown in the upper part of Figure 1.

3.1 Relative Sentence Readability Estimator

Our sentence readability estimator is based on fine-tuning pre-trained masked language models (Devlin et al., 2019). From sentence pairs in parallel corpora for text simplification, we create input sequences of "`[CLS] complex sentence [SEP] simple sentence`" with special tokens indicating the beginning and sentence boundaries. Note that in 50% of the input sequences, the positions of complex and simple sentences are swapped. We train a binary classifier with this dataset to estimate which of two given sentences is more readable.

3.2 Readability Ranking

We rank each sentence in a given set of shuffled sentences by readability using a pairwise comparison method. In other words, the relative sentence readability is estimated for all combinations of two sentences in a given set of sentences, as shown in the bottom right-hand corner of Figure 1. The readability of a sentence is given as an integer, if there is a tie, it depends on the order of input. Finally, we obtain a ranking according to the probability that each sentence is estimated to be more readable.

	Train	Valid	Test
Newsela	385,270	42,323	43,171
CEFR-SP	-	-	17,676

Table 1: Corpus size

4 Experiments

In this section, we experiment with ranking a set of English sentences by readability. Following previous studies of document readability rankings, we evaluated rankings according to four metrics: normalized discounted cumulative gain (NDCG), Spearman's correlation (ρ), Kendall's correlation (τ), and ranking accuracy (RA).

4.1 Experimental Setup

Datasets As shown in Table 1, a relative sentence readability estimator was trained on a training set of $385k$ sentence pairs and a validation set of $42k$ sentence pairs from the parallel corpus for text simplification, Newsela[1] (Xu et al., 2015; Jiang et al., 2020). A set of $43k$ sentence pairs for evaluation was used to construct a set of sentences for readability ranking.[2] We also constructed a set of sentences from the CEFR-SP[3] (Arase et al., 2022), an English corpus with absolute sentence readability. Note that since the CEFR-SP is not a parallel corpus, it is a set of non-synonymous sentences, unlike Newsela. The CEFR-SP has six levels of readability labels for each of the $17k$ sentences, and we randomly selected one sentence at each level to obtain a set

[1] https://github.com/chaojiang06/wiki-auto

[2] Newsela is a parallel corpus consisting of English news articles manually simplified into four levels. In this experiment, sets of synonymous sentences consisting of different simplifications for the same source sentences were ranked in terms of their readability.

[3] https://github.com/yukiar/CEFR-SP

of sentences. Finally, the set of sentences for evaluation from Newsela totals $4,478$ pairs of five level sentences and one from CEFR-SP totals 165 pairs of six level sentences.

Model For our sentence readability estimator, we employed BERT[4] (Devlin et al., 2019) as a pretrained model. We used batch size of 128 sentence pairs, AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 5×10^{-5}. We employed early stopping for fine-tuning with a patience of 3 epochs using a cross-entropy loss in the validation set.

4.2 Baseline models

Baseline unsupervised models We employed two comparative methods of unsupervised sentence readability estimation: define RSRS (Martinc et al., 2021) based on language model scores[5] and methods based on in-context learning of large language models (LLM). The overall system process follows the upper part of Figure 1, as in our method. In addition, RSRS and LLM estimate the readability of each sentence in the set, as shown in the bottom left-hand corner of Figure 1. In this case, RSRS is a floating, and LLM is an integer.

For the LLM-based method, we used LLaMA[5] (Touvron et al., 2023) in two settings, 0-shot and 10-shot. We used the prompts in Figure 2 for experiment, which we modified for sentence readability estimation from the prompts used in a previous study (Wang et al., 2024) working on document readability estimation. 0-shot, in which no examples are presented in the prompt (the "example" portion of Figure 2), and 10-shot, in which 10 examples are presented at each readability level. These examples were randomly selected from valid set from Newsela.

Baseline supervised models We employed two types of baselines for supervised sentence readability estimation: the Pointwise method, which imputes sentences, and the Pairwise method, which imputes sentence pairs. The overall system process follows the upper part of Figure 1, as in our method. These are sentence readability estimation models based on masked language models as in the proposed method, but they were trained using absolute sentence readability labels in the Newsela corpus,[6] unlike the proposed method. The pointwise method is a regression model that estimates the readability of input sentences using masked language models and obtains a readability ranking by the readability of each sentence. This baseline estimates the readability of each sentence in the set, as shown in the bottom left-hand corner of Figure 1. In this case, pointwise is a floating.

System Prompt:
Evaluate the readability of the text using the following eleven levels (reading difficulty):
[score: 2]: Most Easy
[score: 12]: Most Difficult
Based on the provided text examples, assign a readability score to new text and display it in the following format: "[score: X]"
User Input:
Text: {example 1}
[score: 2]
...
Text: {example n}
[score: 12]
New text:
Text: "{}"

Figure 2: Prompt for LLM-based readability estimation.

The pairwise method inputs two sentences as in the proposed method, but unlike the proposed method, it is a regression model that estimates the difference in readability between two given sentences. The pairwise method can provide a binary classification of which sentence is more readable according to whether the output score is positive or negative, resulting in a readability ranking as in the proposed method.

4.3 Results

Readability ranking on Newsela The left side of Table 2 shows the experimental results of readability ranking for a set of synonymous sentences in Newsela[7]. Our method consistently achieved the

[4] `https://huggingface.co/google-bert/bert-base-uncased`

[5] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

[6] There are only a few corpora with sentence readability. So, following previous studies on text simplification (Scarton and Specia, 2018; Nishihara et al., 2019; Yanamoto et al., 2022), the readability of a sentence is defined as the readability of a document containing that sentence. However, but we understand that this is not the best approach.

[7] In this experiment, we use Newsela to enable evaluation, but our method does not use readability labels.

		Newsela (Parallel)				CEFR-SP (Non-Parallel)			
	Supervised	NDCG	ρ	τ	RA	NDCG	ρ	τ	RA
RSRS	-	0.913	0.402	0.341	0.081	0.851	0.082	0.060	0.000
LLM (0-shot)	-	0.888	0.207	0.178	0.041	0.861	0.034	0.027	0.000
Ours	-	**0.985**	**0.865**	**0.799**	**0.421**	**0.958**	**0.749**	**0.619**	**0.048**
Pointwise	✓	0.980	0.841	0.769	0.369	0.949	0.661	0.529	0.012
Pairwise	✓	**0.986**	**0.874**	**0.811**	**0.438**	0.961	0.755	0.621	0.048
LLM (10-shot)	$*$	0.953	0.644	0.550	0.130	**0.967**	**0.764**	**0.636**	**0.073**

Table 2: Experimental results of sentence readability estimation. For each setting, unsupervised and supervised, the highest performance is highlighted in bold. $*$ is a few-shot in-context learning.

best performance among the unsupervised methods in the upper rows. The fact that the pairwise method performed better than the pointwise method among the supervised methods suggests that it is important to consider the relationship between sentences for relative readability estimation. Although the supervised pairwise method showed the best performance, our proposed method in an unsupervised manner also achieved comparable performance. Furthermore, the proposed method outperforms the supervised pointwise method and the LLM-based method in the few-shot setting, revealing its effectiveness.

Readability ranking on CEFR-SP The right side of Table 2 shows the experimental results of readability ranking for a set of non-synonymous sentences in CEFR-SP. Similar to the experimental results on Newsela, the proposed method achieved the best performance among the unsupervised methods in the upper rows. However, experiments with non-synonymous sentence sets showed significantly lower RA overall. In comparison with the supervised methods, the proposed method outperforms the pointwise method and is comparable to the pairwise method, again similar to the experimental results on Newsela. In CEFR-SP, the LLM-based method with the few-shot setting outperformed the other supervised methods, achieving the best performance.

4.4 Analysis

We analyse in detail an experiment on readability ranking for synonymous sentence sets in Newsela.

Is relative sentence readability estimation easier the larger the difference in readability between sentence pairs? → Yes. To clarify this, we append experiments. Table 3 shows the accuracy results of the readability estimation by splitting the sentence pairs in different levels of readability. The results of this analysis show that as the difference in readability increases (more levels of simplification), the accuracy of relative readability estimation improves. As expected, we can conclude that the larger the difference in readability between sentence pairs, the easier the relative readability estimation is. In specific examples, sentence pairs with small differences, such as "Sub-Saharan Africa has benefited from **high** oil and other commodities **prices**, which have started to decline sharply. → Sub-Saharan Africa has benefited from **high prices for** oil and other commodities, which have started to decline sharply.", which is a one-level simplification, have a small difference in readability, and it is difficult to determine the latter sentence is simpler. On the other hand, sentence pairs with large differences, such as "Any artifacts linked to an emperor would bring tremendous pride to Mexico. → Finding remains of those leaders would make Mexico proud.", This is a four-level simplification, has a large difference in readability, and it is easy to determine the latter sentence is simpler. In fact, our method failed to estimate the readability in the top example and succeeded in the bottom one.

Difference in readability	Accuracy
1	0.759
2	0.886
3	0.954
4	0.990

Table 3: Analysis of the impact of differences in readability of sentence pairs on readability estimation.

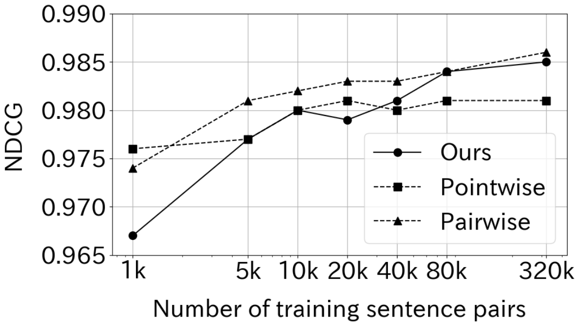

Figure 3: Analysis of the impact of training data size on readability estimation.

How many sentence pairs of parallel corpus for training text simplification make the proposed method effective? → 5k sentences pairs. To clarify this, we append experiments. Figure 3 shows the results of evaluating the quality of the readability ranking (NDCG) while reducing the training data of $385k$ sentence pairs from $320k$ to $1k$ sentence pairs. The results of this analysis show that the text simplification parallel corpus for training our method performs better than the unsupervised sentence readability estimation of RSRS and LLM, NDCG = 0.967 for $1k$ sentence pairs only. As a text simplification parallel corpus of this scale is available in several languages including Japanese, so the method is promising for the multilingual deployment of sentence readability estimation. And if we can prepare a text simplification parallel corpus consisting of $5k$ sentence pairs, to reach comparable performance with supervised sentence readability estimation.

5 Conclusion

In this study, we approach unsupervised sentence readability estimation, which does not use absolute sentence readability data. We train a relative sentence readability estimator that predicts which of two given sentences is more simple, using a text simplification parallel corpus, in our method. Then, we derived a readability ranking for the sentence set by pairwise comparisons. Experimental results in English show that our method outperforms previous unsupervised sentence readability estimation for both synonymous and non-synonymous sentence sets, and achieves performance comparable to supervised methods trained with absolute sentence readability.

Limitations

Although the proposed method was designed with multilingual applications in mind, the experiments in this paper are limited only to English. There is no guarantee that performance consistent with this experiment will be achieved in other languages. As mentioned in Section 1, corpora annotated with sentence readability are scarce, and annotating them is very expensive, therefore, it is not easy to actually experiment with non-English languages.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-Driven Sentence Simplification: Survey and Benchmark. *Computational Linguistics*, 46(1):135–187.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-Based Sentence Difficulty Annotation and Assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.

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, Defence Technical Information Center Document.*

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the Seventh International Conference on Learning Representations.*

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

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

Carolina Scarton and Lucia Specia. 2018. Learning Simplifications for Specific Target Audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 712–718.

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic Assessment of Absolute Sentence Complexity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4096–4102.

Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting Texts by Readability. *Computational Linguistics*, 36(2):203–227.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*.

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.

Ziyang Wang, Sanwoo Lee, Hsiu-Yuan Huang, and Yunfang Wu. 2024. FPT: Feature Prompt Tuning for Few-shot Readability Assessment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 280–295.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Daiki Yanamoto, Tomoki Ikawa, Tomoyuki Kajiwara, Takashi Ninomiya, Satoru Uchida, and Yuki Arase. 2022. Controllable Text Simplification with Deep Reinforcement Learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 398–404.

Victoria Yaneva, Constantin Orăsan, Richard Evans, and Omid Rohanian. 2017. Combining Multiple Corpora for Readability Assessment for People with Cognitive Disabilities. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 121–132.