Promoting Pre-trained LM with Linguistic Features on Automatic Readability Assessment

Shudi Hou, Simin Rao, Yu Xia, Sujian Li
MOE Key Lab of Computational Linguistics, Peking University, Beijing, China
{housd,raosimin,yuxia,lisujian}@pku.edu.cn

Abstract

Automatic readability assessment (ARA) aims at classifying the readability level of a passage automatically. In the past, manually selected linguistic features are used to classify the passages. However, as the use of deep neural network surges, there is less work focusing on these linguistic features. Recently, many works integrate linguistic features with pre-trained language model (PLM) to make up for the information that PLMs are not good at capturing. Despite their initial success, insufficient analysis of the long passage characteristic of ARA has been done before. To further investigate the promotion of linguistic features on PLMs in ARA from the perspective of passage length, with commonly used linguistic features and abundant experiments, we find that: (1) Linguistic features promote PLMs in ARA mainly on long passages. (2) The promotion of the features on PLMs becomes less significant when the dataset size exceeds $\sim$ 750 passages. (3) Our results suggest that Newsela is possibly not suitable for ARA. Our code is available at https://github.com/recorderhou/linguistic-features-in-ARA.

1 Introduction

Readability is proved to be an objective and consistent (Fry, 2002) criterion to level reading materials for language learners. Leveled reading materials are extensively needed, since language learners at different stages of language acquisition need readings at different readability levels to build up their reading skills (Kasule, 2011; Alowais and Ogdol, 2021; Pitcher and Fang, 2007). However, judging and selecting the readability levels of materials need time and professional knowledge, which is quite inefficient compared to the ever-increasing demand. To address the need for automatically assessing the readability level of a given text, Automatic Readability Assessment (ARA) is proposed.

In the early time, experts design formulas (Lennon and Burdick, 2004; Chall and Dale, 1995; Mc Laughlin, 1969; Flesch, 1948) based on the statistics from text such as word length and sentence length. Later, researchers (Feng et al., 2010; McCarthy and Jarvis, 2010; Kate et al., 2010; Vajjala and Meurers, 2012) mine useful morphological, lexical, syntactic and discourse features from text and use them with traditional machine learning models.

Deep learning models such as RNN-based models (Azpiazu and Pera, 2019; Yang et al., 2016) automatically learn dense word embeddings related to the readability of the texts. Recently, the popular pre-trained language models (PLMs) like BERT (Devlin et al., 2019) with their representative dense embeddings are also reported effective (Martinc et al., 2021) on ARA . However, researchers also find handicaps of these deep learning models. Since organizing large-scaled ARA dataset is difficult due to the time and expertise required, datasets used in ARA are relatively small. The insufficiency of data makes it difficult to train a reliable deep learning model (Lee et al., 2021). What's more, as the materials are designed to guide learners step by step, while describing the same thing, the word use, the structure of sentences and the manner of writing the full passages are made stratified intentionally, which is hard to detect for PLMs inclined to semantic information (Martinc et al., 2021; Qiu et al., 2021). For these reasons, some of them incorporate linguistic features with PLMs (Lee et al., 2021; Qiu et al., 2021) and achieve improvements.

Despite their initial success, insufficient analysis of the long passage characteristic of ARA has been done before. We notice that the length of passages in ARA datasets consisting of reading materials can easily go beyond the capacity of PLMs (usually 510 tokens). Specifically, as shown in Fig 1, most ARA datasets have more than 50% passages longer than 510 tokens. Through preliminary experiments (Table 2 last row), we find that such a small dataset is not sufficient to train long-document transform-

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 (Volume 2: Short Papers), pages 430–436
November 20–23, 2022.

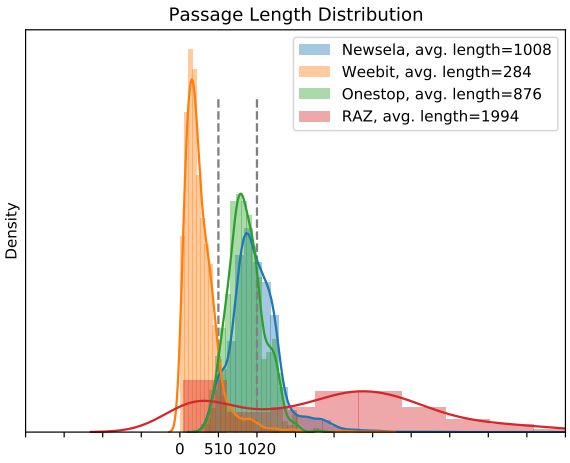

Figure 1: Passage length distribution of 4 datasets.

ers such as CogLTX (Ding et al., 2020) since they usually have more parameters. Besides, splitting passages into shorter pieces and directly congregating them will lose their inner relation, which is sub-optimal for ARA, since characteristics such as the number of theme and the intertextual dependence[1] are important for deciding the readability level. From this point of view, linguistic features extracted from the whole passage actually provide us information from a holistic view, and it can be easily integrated into the models we are using. In this paper, we integrate linguistic features with PLMs and conduct abundant experiments to analyze the effect of linguistic features on ARA from the perspective of passage length. We find that:

- Even with simple linguistic features, the accuracy of PLMs on those small-scaled datasets (OneStop and RAZ) greatly improves by 9% and 22% respectively. Error analysis shows that all of the improvements are on long passages more than 510 tokens.
- The promotion of the features on PLMs becomes less significant when the dataset size exceeds $\sim$ 750 passages.
- Our results suggest that Newsela is possibly not suitable for ARA.

Also, we construct an up-to-date and high-quality dataset called RAZ from RAZ-Kid[2]'s printed leveled books. Though small-scaled, texts from this popular website make our research more practical.

[1]https://www.raz-kids.com/main/ViewPage/name/text-leveling-system/

[2]https://www.raz-kids.com/

Dataset	Long Passages	Rephrase?	#Class	#Passage
Newsela	95.6%	Yes	5	9522
Weebit	10.5%	No	5	3125
OneStop	95.8%	Yes	3	560
RAZ	78.9%	No	3	370

Table 1: Characteristics of 4 datasets. Long passages denote passages with more than 510 tokens.

2 Data Analysis

To analyze the effect of linguistic features as precisely as we could, we select four different datasets namely Weebit (Vajjala and Meurers, 2012), Newsela (Xu et al., 2015), OneStopEnglish (Vajjala and Lučić, 2018) and RAZ. The characteristics of the 4 datasets are listed in Table 1.

Newsela is a text simplification dataset divided into 5 simplification levels. Texts from the hardest level are rephrased 4 times to create other 4 easier levels. Following previous works, we consider each simplification level a readability level.

Weebit is an ARA dataset. Texts from different readability levels focus on different topic. We sample 625 instances each level to construct a balanced dataset.

OneStopEnglish is a relatively small text simplification dataset containing 560 passages. Similar to Newsela, it is also constructed by rephrasing.

RAZ is an ARA dataset constructed by us. We select 370 passages from the RAZ-Kid[2], an online education platform providing lots of leveled eBooks. We manually annotate them with 3 different readability levels according to the readability level criterion[1]. Compared to the above datasets, RAZ contains more text genres, topics and up-to-date vocabulary. More importantly, the average length of RAZ is much longer than the other three datasets, indicating that it is suitable for exploring the effect of linguistic features on long passages.

3 Method

Task Description Given a dataset $\mathcal{D} = \{p_1, p_2, \ldots, p_{|\mathcal{D}|}\}$ with d readability levels $C = \{c_1, c_2, \ldots, c_d\}$. Each passage p_i in dataset $\mathcal{D}$ is mapped to one label in C. It can be regarded as a classification task, a ranking task or an ordinal regression task. We take this task as a classification task for its simplicity.

		Newsela			Weebit			OneStop			RAZ		
		whole	long	short	whole	long	short	whole	long	short	whole	long	short
w/ f_{full}		0.856	0.849	0.965	0.913	0.947	0.908	**0.946**	0.941	1.0	**0.937**	0.911	1.0
w/ $f_{partial}$		0.853	0.851	0.877	0.914	0.957	0.908	0.881	0.847	1.0	0.883	0.847	1.0
w/o feature		**0.876**	0.868	0.982	**0.919**	0.965	0.924	0.863	0.859	1.0	0.766	0.696	1.0
Statistic Model	SVM	0.425	0.561	0.433	0.472	0.482	0.471	0.308	0.309	0.286	0.784	0.8	0.75
	LR	0.711	0.696	0.930	0.599	0.805	0.572	0.75	0.731	1.0	0.784	0.68	1.0
	RF	0.663	0.645	0.930	0.564	0.778	0.536	0.696	0.673	1.0	0.865	0.8	1.0
	NB	0.627	0.607	0.930	0.400	0.638	0.370	0.643	0.615	1.0	0.865	0.8	1.0
Long-Document Model	CogLTX	0.821	0.806	0.975	0.883	0.980	0.870	0.754	0.741	0.936	0.783	0.72	0.91

Table 2: Acc on Newsela, Weebit, OneStopEnglish and RAZ. Results are averaged after three runs for reliability. f_{full} and $f_{partial}$ are defined in Section 3. Long/short denotes the passages longer/shorter than 510 tokens.

Model For each passage $p = [x_1^p, x_2^p, .., x_L^p]$ which has L tokens, we concatenate our extracted linguistic features f_p (see Table 4 for details) and the final hidden state of PLM h_p to form vector $H_p = [h_p, f_p]$. We feed H_p into the classification head of PLM to get the predicted readability level of passage p. Depending on the range of the extracted passage, there are two kinds of features f_p: (1) f_{full} is extracted from the whole passage, which provides a holistic view of the passage; (2) $f_{partial}$ is extracted from the first 510 tokens of p when its length L is greater than 510, which provides the corresponding part of features w.r.t. the segment fed into the PLM. We also report the performance of statistic models using the same linguistic features for comparison.

Implementation Details In our experiments, we use Roberta-base (Liu et al., 2019) as the PLM. While training, we use early stopping based on the accuracy on the dev set. We set the batch size as 8, the max sentence length as 512. We evaluate the model each 50 steps for 100 times. We use AdamW as our optimizer with the learning rate 1e-5 for the PLM encoder and learning rate 1e-3 for the PLM's classification head. The size of train/dev/test set is listed in Tab 3. The linguistic features used in our work are listed in Table 4. We adopt the lexical and syntactic features from (Vajjala and Meurers, 2012) and add some common features from shallow, part-of-speech and discourse aspects. Please refer to our code for more details.

Dataset	train	dev	test
Newsela	7619	952	951
Weebit	2500	313	312
OneStop	448	56	56
RAZ	296	37	37

Table 3: The size of train/dev/test set.

Category	Feature
Shallow Features	Number of Sentences
	Average Sentence Length
	Average Word Difficulty
	Average Word Length
	Number of Uncommon Words
	Number of Unique Words
	Words with 1 to 3 syllables
	Words with 4 syllables
	Words with 5 syllables
	Words with 6 syllables
	Words with more than 7 syllables
	Average number of syllables
POS Features	Number of each POS tags
	POS Divergence
Lexical Features	TTR
	Corrected TTR
	Bi TTR
	Root TTR
	Uber TTR
	Verb Variation-1
	Noun Variation
	Adjective Variation
	Adverb Variation
	Mean Textual Lexical Density
Syntactic Features	Avg Parse Tree Height
	Max Parse Tree Height
	Max Clause Num
	Mean Clause Num
	Max SBAR Num
	Mean SBAR Num
	Max ratio of Dependency Clause
	Mean Ratio of Dependency Clause
Discourse Features	Number of Co-conjection

Table 4: Linguistic features used in our work. The meaning of each feature is detailed in Appendix A.

4 Results and Discussion

4.1 Effect of Linguistic Features: An Overview

In this section, we investigate how linguistic features affect PLMs' performance on ARA. We assume that linguistic features promote PLM in two ways: First, they provide linguistic information that PLM is not good at capturing. Second, they provide information about the segment dropped by PLM, i.e. tokens longer than 510.

To verify our first assumption, we choose $f_{partial}$ as f_p to get H since $f_{partial}$ are the exact corresponding part of features w.r.t. the segment fed into the PLM. Comparing the first and the second row of Table 2, we can see that PLM's performance on RAZ and OneStop improves after adding the features. In Section 4.2, through error analysis, we find that the improvements are all on long passages. The results on Weebit remain almost the same, there are two possible reasons: (1) (Lee et al., 2021) claim that "the max performance (91%) is already achieved on Weebit"; (2) Weebit is 5 to 8 times larger than RAZ and OneStop, such an amount of data is enough for the model to fit well. In Section 4.3, we further investigate the effect of features on different sizes of Weebit and find that features work when we decrease the size of Weebit. The results on Newsela are not as we expected, and we will discuss it in Section 4.4.

To verify our second assumption, we choose f_{full} as f_p to get H since f_{full} provide information about the segment dropped by PLM. Adding these features further improves the PLM's performance on RAZ and OneStop as expected. Specifically, the accuracy of PLMs on these small-scaled datasets greatly improves by 9% and 22% respectively.

4.2 Effect on Long and Short Passages

In order to further analyze on which passages do linguistic features promote PLM, we divide the whole dataset into long and short passages according to whether the passage exceeds 510 tokens. From Fig 2 (right) we can see that the PLM makes no mistake on short passages of RAZ and OneStop. This indicates that the information captured by PLM is enough to classify the short passages even when the dataset is small. From Fig 2 (middle) we can see that $f_{partial}$ reduce the mistakes on long passages without degrading the performance on short passages, and f_{full} further improve the performance greatly, which supports our assumptions. The results on Weebit and Newsela do not match our expections, but they do not conflict with our assumptions. We will discuss them in the following sections.

4.3 Analysis of Dataset Size

As discussed in Section 4.1, the features do not work on Weebit and Newsela. We guess it might be related to the size of dataset since Newsela and Weebit are much larger than RAZ and OneStop (Fig. 1). To analyze the effect of dataset size, we randomly sample 1%, 3%, 5%, 10%, 30%, 50%, 70% of the whole training set of Weebit and Newsela.

Fig. 3 shows that linguistic features significantly improve the PLM's performance on long passages when the dataset size is small (less than 10%). However, as the size exceeds 30% (750 passages)/10% (761 passages) for Weebit/Newsela , the promotion of the linguistic features on PLMs becomes less significant. Although the effect of linguistic features is less significant, we also find out that when the dataset size is between 10% and 50%, the results of PLM with features on both short passages and whole dataset are slightly better than PLM without features. This finding reveals that PLMs cannot learn how to deal with long passages without enough training data, and integrating linguistic features promotes PLMs on long passages. Different from what Lee et al. (2021) find, their simple PLM performs better than our model in the large dataset setting, this is because the features we use are relatively simple. Also, to analyze the effect of features, we do not ensemble traditional statistic models with PLMs, which further restricts the power of features. We think that simple features can already prove our assumptions, so we remain optimistic about the results when more sophisticated features are used and better integration method is applied.

4.4 Text Simplification = ARA?

In this section, we claim that Newsela is possibly not suitable for ARA and consider it an explanation for why the results on Newsela do not meet our expectations. It should be pointed out that ARA focuses on the absolute difficulty of a passage, while text simplification focuses on the relative ranking between different simplified versions of the original passage, which does not ensure one-to-one correspondence between the simplification level and readability level. Measuring the readability level by the Lexile grade just like prior work (Deutsch et al., 2020), we find there is overlap between classes. Specifically, Fig. 4 shows the confusion matrix between the simplification level (SL) and the readability level (RL) on the train set. In order to study to what extent do the overlap affects the performance, we compare the test set accuracy between a non-overlapped set containing 118 passages and a same-sized overlapped set. The results averaged over three runs are 0.646 and 0.453. This indicates

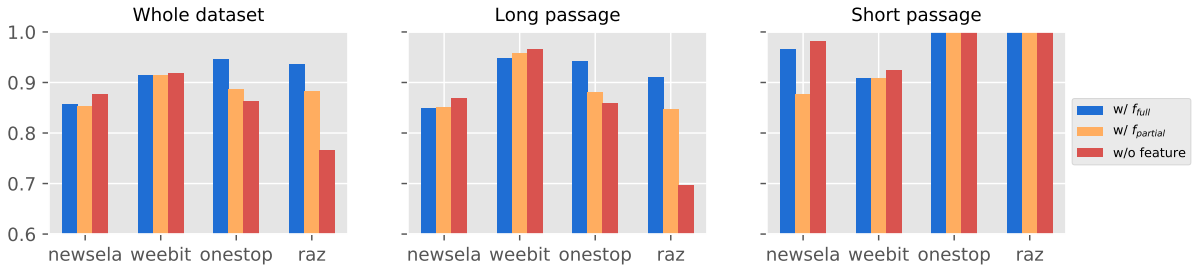

Figure 2: Acc on (left) the whole dataset, (middle) long passages, (right) short passages.

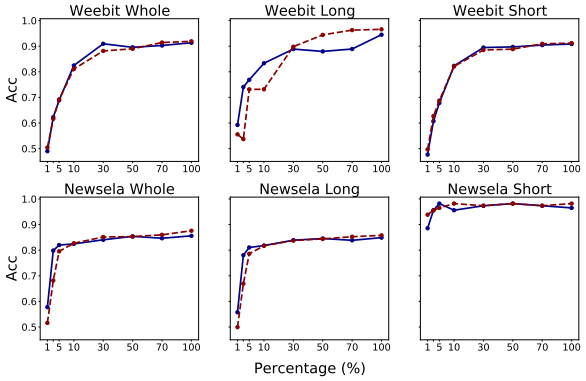

Figure 3: Acc on subsets of Weebit (upper) and Newsela (lower). Blue and red line denotes PLM with f_{full} and without features respectively.

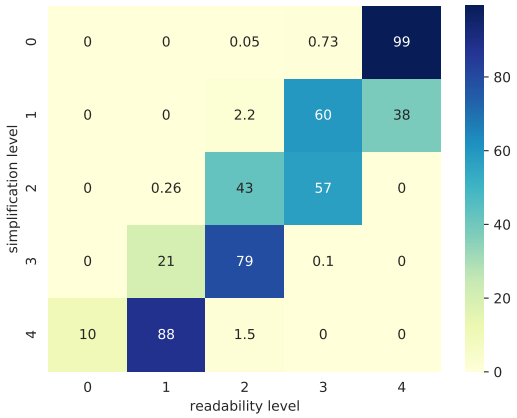

Figure 4: The distribution of Lexile readability level within each simplification level. The Lexile readability level is provided in the Newsela dataset.

that the overlap between classes does confuse the model. Although OneStop is also a text simplification dataset, the three classes are designed to be strictly non-overlapping, thus making OneStop a clean dataset. The insignificant result also indicates that, while integrating linguistic features with PLMs in ARA is effective, it might not be effective for text simplification.

5 Conclusion

In this paper, we investigate how linguistic features promote PLMs on ARA from the perspective of passage length. Firstly, two self-proposed hypotheses are proved: 1. Linguistic features provide linguistic information that PLM is not good at capturing; 2. Linguistic features provide information about the segment dropped by PLM. Secondly, we observe that the promotion of the features on PLMs becomes less significant when the dataset size exceeds ∼750 passages. Thirdly, our results suggest that Newsela dataset is possibly not suitable for ARA.

Acknowledgement

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Natural Science Foundation of China (61876009) and National Social Science Foundation Project of China (21&ZD287). The corresponding author of this paper is Sujian Li.

References

Aisha Alowais and Robin Erric Ogdol. 2021. The effects of leveled reading on second language learners. *International Journal of Research in Education and Science*, 7(4):1281–1299.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39(3):445–459.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

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, Seattle, WA, USA → Online. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284, Beijing, China. Coling 2010 Organizing Committee.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Edward Fry. 2002. Readability versus leveling. *The reading teacher*, 56(3):286–291.

Daniel Kasule. 2011. Textbook readability and esl learners. *Reading & Writing-Journal of the Reading Association of South Africa*, 2(1):63–76.

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554, Beijing, China. Coling 2010 Organizing Committee.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colleen Lennon and Hal Burdick. 2004. The lexile framework as an approach for reading measurement and success. *electronic publication on www. lexile. com*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Brandy Pitcher and Zhihui Fang. 2007. Can we trust levelled texts? an examination of their reliability and quality from a linguistic perspective. *Literacy*, 41(1):43–51.

Xinying Qiu, Yuan Chen, Hanwu Chen, Jian-Yun Nie, Yuming Shen, and Dawei Lu. 2021. Learning syntactic dense embedding with correlation graph for automatic readability assessment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3013–3025, Online. Association for Computational Linguistics.

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.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Category	Features	How to Extract
Shallow Features	Total Number Of Sentences	Count the total number of sentences in a passage
	Average Sentence Length	Average the length of all the sentences in a passage
	Average Word Difficulty	Use IZscore(Balota et al., 2007) to rate the difficulty of a word. If a word's IZscore is bigger than 0, then mark it as a difficult word and we rate this word 1. Otherwise, the word's rate is 0. After rating each word's difficulty, calculate the average of those difficulties.
	Average Word Length	Average the length of all the words in a passage
	Number of Uncommon Words	Count the total number of words that are not in the Dale Chall List
	Number of Unique Words	Count the total number of words that occur in a passage
	Words with 1 to 3 syllables	Count the total number of words with 1-3 syllables
	Words with 4 syllables	Count the total number of words with 4 syllables
	Words with 5 syllables	Count the total number of words with 5 syllables
	Words with 6 syllables	Count the total number of words with 6 syllables
	Words with more than 7 syllables	Count the total number of words with more than 7 syllables
	Average number of syllables	Average each word's syllable number
POS Features	Number of each POS tags	Count the total number of all the POS tags
	POS Divergence	Calculate the KL divergence between sentence POS count distribution and document(Deutsch et al., 2020)
Lexical Features	TTR(Type-Token Ratio)	TTR is the ratio of the number of word types (T) to total number word tokens in a text (N).
	Corrected TTR	$T/\sqrt{2N}$
	Log TTR	$logT/logN$
	Root TTR	$T/\sqrt{N}$
	Uber TTR	$log^2T/logN/T$
	Verb Variation-1	T_{verb}/N_{verb}
	Noun Variation	T_{noun}/N_{lex}
	Adjective Variation	T_{adj}/N_{lex}
	Adverb Variation	T_{adv}/N_{lex}
	Mean Textual Lexical Density	The mean length of sequential word strings in a passage that maintain a given TTR value.(McCarthy and Jarvis, 2010)
Syntactic Features	Avg Parse Tree Height	Calculate the average height of all the constituent trees in a passage.
	Max Parse Tree Height	Calculate the average height of all the constituent trees in a passage
	Max Clause Num	Calculate the max number of clauses in one sentence
	Mean Clause Num	Calculate the average number of clauses in one sentence.
	Max SBAR Num	Calculate the max number of clauses tagged SBAR in one sentence
	Mean SBAR Num	Calculate the average number of clauses tagged SBAR in one sentence.
	Max ratio of Dependency Clause	Calculate the max ratio of dependency clause to all the clause in one sentence
	Mean Ratio of Dependency Clause	Calculate the mean ratio of dependency clause to all the clause in one sentence
Discourse Feature	Number of Co-conjection	Calculate the total number of a co-ordinating conjunction in a passage.

Table 5: The details of linguistic features used in our work. The Dale Chall List could be found at https://readabilityformulas.com/articles/dale-chall-readability-word-list.php

A Linguistic Features Used in Our Work

The meanings of linguistic features are listed in Table 5.