Combining Linguistic and Neural Approaches for Sentence-level Readability Assessment

Anonymous ACL submission

Abstract

Readability assessment, which predicts how difficult it is for the reader to understand a text, has mostly been performed at the passage level. There has been recent interest in sentence-level assessment, given its applications in downstream tasks such as text simplification and language exercise generation. While research has shown that neural approaches can improve the assessment accuracy on the passage levels, they have yet to be applied on the sentence level. This paper evaluates a neural model on readability assessment at the sentence level, and shows that its performance can be further improved by integrating linguistic features. On a dataset of 13,809 Chinese sentences ranging from Grade 3 to 12, we obtained the strongest performance with a model that combines BERT and hand-crafted features, which has previously been applied on identifying propaganda in news (Kaas et al., 2020). The rest of the paper is organized as follows. After a review of previous work (Section 2), we present our dataset (Section 3) and approach (Section 4), and then discuss experimental results (Section 5).

1 Introduction

Text readability is defined as the cognitive load of a reader to comprehend a given text (Martinc et al., 2021). Most research on readability assessment, has focused on estimating the difficulty level of a passage (Azpiazu and Pera, 2019; Lee et al., 2021) rather than on individual sentences. However, sentence-level readability assessment can contribute to many downstream tasks related to natural language generation, such as generation of example sentences and language exercises (Pilán et al., 2014), as well as text simplification (Gărbacea et al., 2021), by assisting editors to identify complicated sentences.

Since neural approaches have been shown to excel in a large variety of natural language processing tasks, they have also been applied on readability assessment of passages (Filighera et al., 2019; Tseng et al., 2019; Deutsch et al., 2020; Martinc et al., 2021; Lee et al., 2021). There has been less research effort, however, on applying them at the sentence level. To the best of our knowledge, there is only one reported evaluation of sentence-level neural readability assessment (Schicchi et al., 2020).

This paper evaluates a neural model on readability assessment at the sentence level, and shows that its performance can be further improved by integrating linguistic features. On a dataset of 13,809 Chinese sentences ranging from Grade 3 to 12, we obtained the strongest performance with a model that combines BERT and hand-crafted features, which has previously been applied on identifying propaganda in news (Kaas et al., 2020). The rest of the paper is organized as follows. After a review of previous work (Section 2), we present our dataset (Section 3) and approach (Section 4), and then discuss experimental results (Section 5).

2 Previous work

2.1 Linguistic features and neural approaches for readability

Readability formulas (Kincaid et al., 1975) and traditional approaches for readability assessment have mostly made use of linguistic features and language models (Collins-Thompson, 2008; Sung et al., 2015). More recent studies have shown that neural approaches can improve assessment performance in various languages (Azpiazu and Pera, 2019; Martinc et al., 2021). An active area of research is to investigate how to incorporate linguistic features into neural models. Some studies on passage-level assessment indicated only marginal improvement (Filighera et al., 2019) or no effect (Deutsch et al., 2020), others reported significant improvement by combining Random Forest and RoBERTa (Lee et al., 2021). However, the effect of combining linguistic features and neural approaches has not yet been investigated at the sentence level.

2.2 Sentence-level readability assessment

A naive application of passage-level readability assessment models on sentences would lead to a sub-
While there are publicly available datasets for sentence-level readability with a binary distinction, such as Wikipedia and Simple Wikipedia (Vajjala and Meurers, 2014), we are not aware of any that is annotated on multiple levels of difficulty. Since many readability datasets are available at the passage level, a possible methodology is to assigning the grade of the passage to all sentences in that passage (Pilán et al., 2014). Substantial noise could be introduced to both the training process and evaluation results, however, as a text can contain easier or more difficult sentences.

We constructed our dataset out of 109 graded texts, ranging from Grade 3 to 12 and covering a variety of genres and topics, taken from a corpus of Chinese-language textbooks used in Mainland China (Lee et al., 2020). There are a total of 13,809 sentences.

A previous study calculated the grade of a sentence based on grammar points in the sentence and the grade of the passage (Lu et al., 2020). However, the difficulty of a sentence could be influenced by many other factors, and native speakers’ rating was found to be more reliable than other automatic methods such as reading time or translation time (Kroeger and Sun, 2005). We therefore recruited two native speakers of Mandarin to perform the annotation. One was a certified Chinese-language teacher with over 26 years of experience, and the other was an undergraduate student majoring in linguistics. The annotators found it subjective and challenging to specify the grade level of each sentence. They were instead asked to read all passages of the same grade, and then identify sentences in each passage to be either easier, harder, or at the expected level of difficulty relative to other sentences.

Table 1 shows three example sentences taken from a Grade 3 passage. The top sentence was judged to be typical, with multiple subjects from tangmu ‘Thomas’ to hua ‘speech’ and shengyin ‘voice’. The middle sentence has a simpler structure, with only subject shaonian ‘youth’. The bottom sentence is more difficult to read due to the complex modifiers for the noun xinyin ‘trust’ and for the noun xiyou ‘joy’. The agreement percentage between the two annotators is 77.22%. They achieved a Cohen’s Kappa of 0.55, corresponding to a moderate level of agreement (Landis and Koch, 1977). They reconciled the label differences through discussion to reach the final annotation.

On average, 8.72% of the sentences are labeled as more difficult than expected, and 32.37% less difficult. The number of easier sentences exceeds the number of harder sentences in all grades with the exception of Grade 3. As shown in Figure 1, in general, the higher the grade, the smaller the proportion of sentences labeled as “harder”. The proportion of sentences labeled as “easier” varies from 37.95% to 10.24%. Statistics on our corpus can be found Table 2.\footnote{This dataset is not publicly available.} \footnote{All data and code will be publicly released for research purposes upon publication of this paper.}
Sentence Difficulty

汤姆开头有点吞吞吐吐，渐渐地，话越来越多，声音也越来越大，越来越自然了。
‘Thomas stuttered at first, but gradually his speech got longer, his voice grew louder and more natural.’

少年连连摆手，用不太标准的中国话说：“不，不要钱。”
‘The youth waved his hands and said with an accent, “No, I don’t want money.”’

在那里，我们得到的是人与人之间的信任和被信任的喜悦。
‘There, what we get is the trust between people and the joy of being trusted.’

Table 1: Example sentences taken from a Grade 3 passage, annotated with their difficulty relative to Grade 3.

<table>
<thead>
<tr>
<th>Grade</th>
<th># sent</th>
<th>Grade</th>
<th># sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>859</td>
<td>8</td>
<td>707</td>
</tr>
<tr>
<td>4</td>
<td>111</td>
<td>9</td>
<td>895</td>
</tr>
<tr>
<td>5</td>
<td>320</td>
<td>10</td>
<td>1901</td>
</tr>
<tr>
<td>6</td>
<td>347</td>
<td>11</td>
<td>892</td>
</tr>
<tr>
<td>7</td>
<td>1372</td>
<td>12</td>
<td>1632</td>
</tr>
</tbody>
</table>

Table 2: Number of sentences classified to be at the expected difficulty at each grade

<table>
<thead>
<tr>
<th>Model</th>
<th>Annotation</th>
<th>10-way</th>
<th>4-way</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>n/a</td>
<td>21.05%</td>
<td>40.81%</td>
</tr>
<tr>
<td>LR</td>
<td>All</td>
<td>17.19%</td>
<td>41.67%</td>
</tr>
<tr>
<td>RF</td>
<td>All</td>
<td>24.78%</td>
<td>49.53%</td>
</tr>
<tr>
<td>XGB</td>
<td>All</td>
<td>18.77%</td>
<td>51.99%</td>
</tr>
<tr>
<td>BERT</td>
<td>All</td>
<td>35.73%</td>
<td>50.20%</td>
</tr>
<tr>
<td></td>
<td>Nil</td>
<td>37.02%</td>
<td>47.13%</td>
</tr>
<tr>
<td></td>
<td>Exp</td>
<td>37.05%</td>
<td>47.15%</td>
</tr>
<tr>
<td>RF+BERT</td>
<td>All</td>
<td>36.85%</td>
<td>49.37%</td>
</tr>
<tr>
<td>Aug. BERT</td>
<td>All</td>
<td>38.55%</td>
<td>58.28%</td>
</tr>
</tbody>
</table>

Table 3: Readability assessment performance for various models and annotation methods during training

4 Approach

We trained a number of classifiers by traditional machine learning, a neural classifier and a modified neural architecture than combines linguistic features.

4.1 Baseline classifiers

We trained classifiers with Logistic Regression (LR), Random Forests (RF) and Xgboost (XGB), using the implementation of scikit-learn (Pedregosa et al., 2011). Similar to Lu et al. (2020), we trained these classifiers with bag-of-word features eatures as well as a set of 41 linguistic features from raw sentences, lexical, syntactic and discourse features. The complete description of linguistic features are attached in the Appendix.

4.2 Neural classifier

BERT (Devlin et al., 2019) has achieved the state-of-the-art performance in many natural language processing tasks. We used the pre-trained model bert-base-chinese that contains 12 layers, 768 hidden units and 12 heads. We fine-tuned this model on our dataset (Section 3) into a 10-way classifier for sentence-level readability assessment. This approach is similar to one taken by Tseng et al. (2019) for passage-level readability assessment.

4.3 Integration of linguistic features

RF+BERT. Following Deutsch et al. (Deutsch et al., 2020) and Lee et al. (Lee et al., 2021), we obtained the grade predicted by BERT (Section 4.2) and added it as an additional feature to Random Forest (RF) (Section 4.1), which turned out to be the best-performing baseline classifier (Section 5.4).

Augmented BERT. We used the modified BERT model architecture used by Kaas et al. (2020), which was shown to be effective in propaganda detection in news articles. In the first component, linguistic features (Section 4.1) are fed into the one linear layer, and then passed to one dense layer and ReLU layer, including dropout parameter. In the second component, the outputs of BERT hidden layer are passed to one dense layer and ReLU layer, also including dropout parameter. Finally, these two vectors are concatenated to a fully-connected layer and softmax layer to predict the final sentence readability level.

3 We use the Adam algorithm (Kingma and Ba, 2015) for optimization. The epoch for each training is 20, and set the maximum word embedding size as 128.

4 The design of the model was in turn inspired by Zhang and Li.
5 Experiments

5.1 Annotation methods for training

We evaluated three annotation methods for training:

**Nil** Assign the grade of the passage to all sentences in the training set. This method avoids the need to annotate the sentence level in return for some noise in the training labels.

**All** A sentence labeled as “expected” is assigned the grade of its passage; a sentence labeled as “easier” is assigned one grade below the grade of its passage; and a sentence labeled as “harder” is assigned one grade above. For example, in Table 1, the “easier” sentence is considered Grade 2 and the “harder” sentence Grade 4. This method is designed to reduce the noise in the training labels.

**Exp** Use only sentences labeled as “expected”, and assign the grade of the passage to these sentences. This method results in a smaller training data size but removes noise in the training labels.

5.2 Testing

To ensure the correctness of the gold label, we tested only those sentences labeled as “expected”. We used stratified ten-fold cross validation in all experiments, with 8 folds as training set, 1 fold as dev set and 1 fold as test set. The hyperparameters learning rate, dropout and batch size are tuned on the dev set, and found the best parameters at about learning rate as $1e^{-5}$, dropout as 0.2 and batch size as 64. All sentences from the same text are placed in the same fold, so that the entities mentioned in the test sentences would not be seen during training.

5.3 Metrics

We report three metrics: 10-way classification accuracy on predicting the grade of the sentence as annotated in the corpus; adjacency accuracy, i.e. allowing the prediction to deviate by one grade from the gold; and 4-way accuracy, by merging Grades 1-3, Grades 4-6, Grades 7-9 and Grades 10-12 into four difficulty levels.

5.4 Results

**Neural vs. statistical classifiers.** As shown in Table 3, among the baseline classifiers, XGB yielded the highest accuracy on both 4-way classification (47.34%) and 10-way adjacency classification (51.99%). In terms of 10-way classification, Random Forest (RF) performed best at 24.78%. The BERT model substantially improved upon the performance of RF on 10-way classification to 35.73%. However, its performance is only slightly better than XGB in 4-way classification, and slightly worse in 10-way adjacency classification.

**Annotation method.** The Exp method — i.e., using only the “expected” sentences and therefore have the precise grade level — led to the best accuracy for 10-way classification (37.05%). This is likely due to a lower level of noise in the training data. For 10-way adjacent classification and 4-way classification, The All method performed best, which suggests that a larger but noisier dataset is beneficial for more coarse-grained tasks.

**Integration of linguistic features.** The RF+BERT model outperforms the BERT model only slightly for 10-way classification and 4-way classification, and performed slightly worse in adjacency classification. The Augmented BERT model achieved the best result in all three metrics, significantly outperformed the BERT model on 10-way classification (38.55%) and adjacency classification (58.28%), as well as 4-way classification (54.97%). These results suggest that this modified model can better combine the insights from the linguistic features and the neural network in comparison to the RF+BERT model, which simply includes the BERT model’s prediction into the RF classifier.

6 Conclusion

We have presented the first study on integrating linguistic features into a neural model for sentence-level readability assessment. Our contribution is three-fold. First, we contribute a dataset of over 13K Chinese sentences annotated with their difficulty at the sentence level. Second, we investigated the utility of sentence-level annotation, showing that the “easier” and “harder” labels relative to passage difficulty can improve performance, especially for more coarse-grained assessment. Finally, we evaluated an augmented BERT model that integrates linguistic features, and demonstrated that it outperforms the simple integrated models used in previous studies.

\[^5\text{At } p = 1.1 \cdot e^{-4} \text{ for 10-way Acc., } p = 6.1 \cdot e^{-50} \text{ for 4-way Acc. according to McNemar’s Test.}\]
References


A Appendix: Instructions to annotators

We recruited two annotators, who accepted to work on a volunteer basis, under the project Anonymous at XXX University. The project has been approved by a institutional ethics review board. The annotators were given an Excel file with one sentence per row, and were asked to mark each sentence as 0 (“expected” level of difficulty), +1 (“harder”) or -1 (“easier”).

B Appendix: List of linguistic features

Raw Sentences
1. Number of words per sentence
2. Number of characters per sentence

Lexical Features
3. Average number of characters per word per sentence
4. Number of two-character words per sentence
5. Percentage of two-character words per sentence
6. Number of three-character words per sentence
7. Percentage of three-character words per sentence
8. Number of four-character words per sentence
9. Percentage of four-character words per sentence
10. Number of five-up-character words per sentence
11. Percentage of five-up-character words per sentence
12. Percentage of level1 tokens per sentence
13. Percentage of level2 tokens per sentence
14. Percentage of level3 tokens per sentence
15. Percentage of level4 tokens per sentence
16. Percentage of level5 tokens per sentence
17. Percentage of level6 tokens per sentence
18. Percentage of level7 tokens per sentence

Syntactic Features
19. Percentage of adjectives per sentence
20. Number of adjectives per sentence
21. Percentage of verbs per sentence
22. Number of verbs per sentence
23. Percentage of nouns per sentence
24. Number of nouns per sentence
25. Percentage of adverbs per sentence
26. Number of adverbs per sentence
27. Total number of words tagged as conjunctions
28. Percentage of conjunctions per sentence
29. Total number of words tagged as pronouns
30. Percentage of pronouns per sentence
31. Total number of prepositions per sentence
32. Percentage of prepositions per sentence

Discourse Features
33. Height of parse tree per sentence
34. Total number of noun phrases per sentence
35. Total number of verbal phrases per sentence
36. Total number of prepositional phrases per sentence
37. Total number of dependency distances per sentence
38. Average number of dependency distances per sentence
39. Total number of entities per sentence
40. Percentage of entities per sentence
41. Number of Not-Entity nouns per sentence