Method	Word Acc	Word Corr	Document Acc	Document Corr
HAN	-	-	0.367	0.498
LR (regr)	0.409	0.534	0.480	0.657
LR (cls+m)	0.440	0.514	0.765	0.723
LR (cls+w)	0.440	0.540	0.765	0.880
GBDT	0.432	0.376	0.765	0.833
GCN (regr)	0.434	0.579	0.643	0.849
GCN (cls+m)	**0.476**	0.536	**0.796**	0.878
GCN (cls+w)	**0.476**	**0.592**	**0.796**	**0.891**

Table 2: Difficulty estimation results in accuracy (Acc) and correlation (Corr) on classification outputs converted to continuous values by taking the max (cls+m) or weighted sum (cls+w) and regression (regr) variants for the logistic regression (LR) and GCN.

Evaluation We use accuracy and Spearman's rank correlation as the metrics. When calculating the correlation for a classification model, we convert the discrete outputs into continuous values in two ways: (1) convert the CEFR label with the maximum probability into corresponding β in Section 2, (cls+m), or (2) take a sum of all β in six labels weighted by their probabilities (cls+w).

4.1 Results

Table 2 shows the test accuracy and correlation results. GCNs show increase in both document accuracy and word accuracy compared to the baseline. We infer that this is because GCN is good at capturing the relationship between words and documents. For example, the labeled training documents include an A1 document and that contains the word "bicycle," and the difficulty label of the document is explicitly propagated to the "bicycle" word node, whereas the logistic regression baseline mistakenly predicts as A2-level, since it relies solely on the input features to capture its similarities.

4.2 Ablation Study on Features

Table 3 shows the ablation study on the features explained in Section 4. By comparing Table 2 and Table 3, which are experimented on the same datasets, GCN without using any traditional or embedding features ("None") shows comparative results to some baselines, especially on word-level accuracy. Therefore, the structure of the word-document graph provides effective and complementary signal for readability estimation.

Overall, the BERT embedding is a powerful feature for predicting document readability on Cambridge Readabilty Dataset. Ablating the BERT embeddings (Table 3) significantly decreases the document accuracy (-0.112) which is consistent with the previous work (Martinc et al., 2019; Deutsch et al., 2020) that BERT being one of the best-performing method for predicting document readability on one of the datasets they used, and HAN performing relatively low due to not using the BERT embeddings.

Features	Word Acc	Word Corr	Document Acc	Document Corr
All	0.476	0.592	0.796	0.891
−word freq.	0.476	0.591	0.796	0.899
−doc length	0.481	0.601	0.796	0.890
−GloVe	0.463	0.545	0.714	0.878
−BERT	0.450	0.547	0.684	0.830
None	0.440	0.436	0.520	0.669

Table 3: Ablation study on the features used. "None" is when applying GCN without any features ($X = I$ i.e., one-hot encoding per node), which solely relies on the word-document structure of the graph.

4.3 Training on Less Labeled Data

To analyze whether GCN is robust when training dataset is small, we compare the baseline and GCN by varying the amount of labeled training data. In Figure 3, we observe consistent improvement in GCN over the baseline especially in word accuracy. This outcome suggests that the performance of GCN stays robust even with smaller training data by exploiting the signals gained from the recursive word-document relationship and their structure. Another trend observed in Figure 3 is the larger gap in word accuracy compared to document accuracy when the training data is small likely due to GCN explicitly using context given by word-document edges.

5 Conclusion

In this paper, we proposed a GCN-based method to jointly estimate the readability on both words and documents. We experimentally showed that GCN achieves higher accuracy by capturing the recursive difficulty relationship between words and documents, even when using a smaller amount of labeled data. GCNs are a versatile framework that allows inclusion of diverse types of nodes, such as

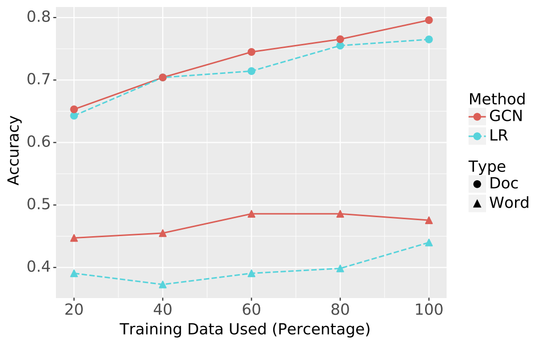

Figure 3: Word and document accuracy with different amount of training data used.

subwords, paragraphs, and even grammatical concepts. We leave this investigation as future work.

Acknowledgements

The authors would like to thank Adam Wiemerslage, Michael J. Paul, and anonymous reviewers for their detailed and constructive feedback. We also thank Kathleen Hall for her help with annotation.

References

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1.

Annette Capel. 2012. Completing the english vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Brent Culligan. 2015. A comparison of three test formats to assess word difficulty. *Language Testing*, 32(4):503–520.

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*.

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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment.

Thomas François, Nùria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Language Resources and Evaluation Conference*.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Language Resources and Evaluation Conference*.

Filip Ginter, Jan Hajič, Juhani Luotolahti, Milan Straka, and Daniel Zeman. 2017. CoNLL 2017 shared task - automatically annotated raw texts and word embeddings. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations*.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Frederic M. Lord. 1980. *Applications of Item Response Theory To Practical Testing Problems*. Lawrence Erlbaum Associates.

Matej Martinc, Senja Pollak, and Marko Robnik-Sikonja. 2019. Supervised and unsupervised neural approaches to text readability. *CoRR*, abs/1907.11779.

Masashi Negishi, Tomoko Takada, and Yukio Tono. 2013. A progress report on the development of the CEFR-J. In *Exploring language frameworks: Proceedings of the ALTE Kraków Conference*, pages 135–163.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

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 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*.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Shuhan Wang and Erik Andersen. 2016. Grammatical templates: Improving text difficulty evaluation for language learners.

Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. 2019. Simplifying graph convolutional networks. In *Proceedings of the International Conference of Machine Learning*.

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*.

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

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Association for the Advancement of Artificial Intelligence*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

A Hyperparameter Details

We conduct random hyperparameter search with 200 samples in the following ranges: $\alpha \in \{0.1, 0.2, ..., 0.9\}$, the learning rate from $\{1, 2, 5, 10, 20, 50, 100\} \times 10^{-4}$, dropout probability from $\{0.1, 0.2, ..., 0.5\}$, the number of epochs from $\{250, 500, 1000, 1500, 2000\}$, the number of hidden units $h_n \in \{32, 64, 128, 256, 512, 1024\}$, the number of hidden layers from $\{1, 2, 3\}$, and the PMI window width from $\{\text{disabled}, 5, 10, 15, 20\}$.

We now describe the selected best combination of hyperparameters for each setting. For GCN in the classification setting, the selected hyperparameters for document difficulty estimation are:

- α: 0.3
- Learning rate: $5 \cdot 10^{-4}$
- Dropout probability: 0.5
- The number of epochs: 500
- The number of hidden units h_n: 512
- The number of hidden layers N: 2
- PMI window width: 5

and for word difficulty estimation, the selected hyperparameters are:

- α: 0.2
- Learning rate: $5 \cdot 10^{-3}$
- Dropout probability: 0.2
- The number of epochs: 250
- The number of hidden units h_n: 64
- The number of hidden layers N: 1
- PMI window width: disabled

For GCN in the regression setting, the selected hyperparameters for document difficulty estimation are:

- α: 0.4
- Learning rate: $2 \cdot 10^{-4}$
- Dropout probability: 0.3
- The number of epochs: 1500
- The number of hidden units h_n: 128
- The number of hidden layers N: 2
- PMI window width: 5

and for word difficulty estimation, the selected hyperparameters are:

- α: 0.2
- Learning rate: $1 \cdot 10^{-3}$
- Dropout probability: 0.1
- The number of epochs: 500
- The number of hidden units h_n: 512
- The number of hidden layers N: 2
- PMI window width: disabled