Hybrid Models for Sentence Readability Assessment

Fengkai Liu, John S. Y. Lee
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
fengkaliu3-c@my.cityu.edu.hk, jsylee@cityu.edu.hk

Abstract

Automatic readability assessment (ARA) predicts how difficult it is for the reader to understand a text. While ARA has traditionally been performed at the passage level, there has been increasing interest in ARA at the sentence level, given its applications in downstream tasks such as text simplification and language exercise generation. Recent research has suggested the effectiveness of hybrid approaches for ARA, but they have yet to be applied on the sentence level. We present the first study that compares neural and hybrid models for sentence-level ARA. We conducted experiments on graded sentences from the Wall Street Journal (WSJ) and a dataset derived from the OneStopEnglish corpus. Experimental results show that both neural and hybrid models outperform traditional classifiers trained on linguistic features. Hybrid models obtained the best accuracy on both datasets, surpassing the previous best result reported on the WSJ dataset by almost 13% absolute.

1 Introduction

Text readability is defined as the cognitive load of a reader to comprehend a text (Martinc et al., 2021). Research on automatic readability assessment (ARA) has traditionally aimed at passages (Azpiazu and Pera, 2019), e.g., labeling a passage with its difficulty level.

There has been growing interest in assessing the difficulty of individual sentences (Štajner et al., 2017; Brunato et al., 2018; Lu et al., 2020; Schicchi et al., 2020), given its application in various downstream tasks in natural language processing (NLP). It is essential to generation tasks that are sensitive to language difficulty, such as pedagogical material and exercises (Pilán et al., 2014). It also facilitates explainable text simplification (Gârbacea et al., 2021) by identifying which sentences require simplification. Sentence-level ARA is a task in its own right since a substantial drop in performance has been observed when passage-level ARA models are applied on individual sentences (Kilgarriff et al., 2008; Pilán et al., 2016).

Similar to many other NLP tasks, passage-level ARA has benefited from the advent of neural approaches (Filighera et al., 2019; Tseng et al., 2019; Martinc et al., 2021). Recent research has also applied 'hybrid' models, which leverage both linguistically motivated features and neural models (Deutsch et al., 2020; Lee et al., 2021; Lim et al., 2022). For sentence-level ARA, although neural models have been evaluated (Schicchi et al., 2020; Arase et al., 2022), there has not been any attempt to integrate linguistic features.

This paper applies neural models and hybrid models on sentence-level ARA and compares their performance with a non-neural classifier trained on linguistic features. To our knowledge, this is the first study on hybrid models for sentence-level ARA. Experimental results show that a hybrid model offers the best performance, and surpasses the previous best result reported on the Wall Street Journal dataset (Brunato et al., 2018). [1]

2 Previous work

2.1 Neural and hybrid approaches

Readability formulas (Kincaid et al., 1975) and traditional approaches for readability assessment have mostly relied on one-hot linguistic features and language models (Collins-Thompson, 2008; Sung et al., 2015). More recent studies have shown that neural approaches can improve assessment performance (Azpiazu and Pera, 2019; Martinc et al., 2021). An active area of ARA research is to investigate how to incorporate linguistic features into neural models. On passage-level assessment, some studies observed no effect (Deutsch et al., 2020) or only marginal improvement (Filighera et al., 2019)

[1] All data and code are publicly released at https://github.com/ffliu6/Hybrid4SentenceARA.

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 448–454

from linguistic features, while others reported significant improvement, e.g. by combining Random Forest and RoBERTa (Lee et al., 2021), and concatenating linguistic features with sentence embeddings from BERT hidden layers (Imperial, 2021). However, there has not yet been any study on hybrid models on sentence-level ARA.

2.2 Sentence readability assessment

Most previous research on sentence readability pursued binary classification or pairwise difficulty prediction (Ambati et al., 2016; Schumacher et al., 2016). An algorithm combining rule-based and statistical classifiers yielded 71% accuracy on binary classification of texts for learning Swedish as a foreign language (Pilán et al., 2014). Statistical classifiers achieved 66% accuracy on an English dataset based on Wikipedia and Simple Wikipedia (Vajjala and Meurers, 2014) and between 78.9% and 83.7% on an Italian dataset (Dell'Orletta et al., 2014).

There have also been a few studies on sentence-level ARA involving multi-way classifiers trained with traditional machine learning methods. Brunato et al. (2018) developed an SVM linear regression model with a variety of surface, morphological and syntactic features. The model achieved 59.1% and 60% accuracy on an Italian and an English dataset of sentences graded on a 7-point scale. Sentence length and nominal modification were found to correlate significantly with sentence difficulty. A Bayesian Ridge Regression Model, trained on a variety of linguistic features including syntax, lexical, morphology and cohesion, has been shown to achieve high correlation with human judgment on German sentence difficulty (Weiss and Meurers, 2022). A classifier has also been trained on features derived from the phrase complexity level of n-grams (Štajner et al., 2017). It attained 0.66 weighted F-score on an English dataset on a 5-point scale. A classifier for Chinese sentences, based on vocabulary and grammar points, reached 31.92% accuracy on 10-way classification (Lu et al., 2020).

Two studies have applied neural models on sentence-level ARA. Schicchi et al. (2020) showed that an RNN-based architecture outperformed Vec2Read (Mikolov et al., 2013). Arase et al. (2022) found that the BERT-base model outperformed traditional machine learning classifiers on their annotated CEFR-based sentence difficulty dataset. However, they did not attempt to incorporate any linguistic features. This paper aims to fill in this gap with a comparison of neural models, hybrid models and traditional classifiers.

3 Data

We used the following two datasets in our experiments. Detailed statistics are shown in Table 3 and Table 4 (see Appendix A).

3.1 Wall Street Journal (WSJ)

This corpus (Brunato et al., 2018) consists of 1,200 sentences drawn from the Wall Street Journal (Nivre et al., 2007) and graded on a difficult scale from 1 to 7. Each sentence was rated by 20 native speakers on a difficult scale from 1 ("very easy") to 7 ("very difficult"). Our evaluation is based on the set of 650 sentences whose grade was agreed upon by at least 14 of the 20 annotators. While it is possible to restrict the evaluation to sentences with an even higher rate of agreement, it would lead to a substantially smaller dataset, whose size is already much smaller than other datasets. [2]

3.2 OneStopEnglish (OSE)

This corpus (Vajjala and Lučić, 2018) consists of aligned texts graded at three reading grades: beginner, intermediate, and advanced. Each of the 189 texts has three versions corresponding to these grades, with a total of 19,904 sentences in the 567 texts. [3]

Instead of assigning the grade of the text to all sentences in that text (Pilán et al., 2014), we determined the difficulty of each individual sentence based on the human revision. Among the sentences in intermediate texts, 10.21% appear verbatim in the beginner version; among those in the advanced texts, 18.76% appear verbatim in one of the lower versions. These sentences are labeled with the lowest grade at which they appear. All other sentences are labeled with the grade of the text — the fact the human editors revised them implies that their grade could not be lower.

4 Approach

4.1 Baseline: Linguistic Model

We used the scikit-learn implementation of Random Forests (RF) and XGBoost (XGB) (Pedregosa

[2] No sentence in this subset was graded at 6 or 7.

[3] Sentence segmentation was performed with NLTK (Bird et al., 2009).

et al., 2011). We extracted 255 linguistic features with LingFeat[4] for each sentence. We performed feature selection with the Variance Threshold in scikit-learn on the dev set.[5] Similar to Lu et al. (2020), we trained these classifiers with linguistic features as well as bag-of-word features.

4.2 Baseline: Neural Model

Transformer-based neural models have achieved impressive performance in many natural language processing tasks.

We fine-tuned BERT (Devlin et al., 2019), BART (Lewis et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020) on our datasets (Section 3) into an ARA classifier[6] , using the pre-trained versions released by Huggingface (Wolf et al., 2019). We used the base versions of all of the above, as well as the large versions of BART, RoBERTa and ELECTRA.

4.3 Hybrid Models

We implemented three hybrid models. The following model incorporates linguistic features into a neural model:

Concatenated Model Similar to Song et al. (2021), the input to model consists of the input sentence $w_1 w_2 \dots w_n$ concatenated with the linguistic features $f_1, f_2 \dots f_n$, in the format "`[CLS]` $w_1 w_2 \dots w_n$ `[SEP]` $f_1 f_2 \dots f_n$".

The following two models wrap the linguistic features and neural model output in a non-neural statistical classifier:

Hard Label Following Deutsch et al. (2020), the grade of the sentence, as predicted by the Neural Model (Section 4.2), serves as an additional feature in the statistical classifier (Section 4.1).

Soft Labels Following Lee et al. (2021), the probability of each grade, as predicted by the Neural Model (Section 4.2), serve as additional features alongside the linguistic features in the statistical classifier (Section 4.1).

[4] https://github.com/brucewlee/lingfeat

[5] The threshold set to 0.8.

[6] We used the Adam algorithm (Kingma and Ba, 2015) for optimization. The epoch for each training is 10, and set the maximum word embedding size as 128.

5 Experiments

5.1 Set-up

We report results in terms of accuracy (Acc.), F1-score, Precision, Recall and QWK scores.

We used stratified ten-fold cross validation in WSJ and OSE experiments, with a 8:1:1 split for training, development and testing.[7] For the OSE dataset, all sentences from the same text are placed in the same fold, so that the entities and topics mentioned in the test sentences would not be seen during training.

5.2 Results

Linguistic Model. XGBoost (XGB) outperformed Random Forest (RF) and Linear Regression (LR) on all datasets. On OSE and WSJ, it achieved 0.451 and 0.618 accuracy, respectively, compared to 0.412 and 0.551 for RF, and 0.374 and 0.413 for LR. We will therefore present results based on XGB in the remainder of this section.

Neural Model. Table 1 presents the performance of neural models on the WSJ and OSE datasets. On the WSJ dataset, RoBERTa obtained the best performance among base versions, at a 0.668 accuracy. Large models were found to outperform base versions on the WSJ dataset, in which BART-large produced the highest accuracy at 0.679. On the OSE dataset, BART obtained the best performance among base versions, at a 0.571 accuracy. Large models were also found to outperform base versions on the OSE dataset, in which BART-large produced the highest accuracy at 0.571. Generally, BART-large model achieved the best performance on all datasets, at 0.679 and 0.571 accuracy for the WSJ and OSE datasets, respectively. We will therefore use its predictions for hybrid models.

The results for OSE and WSJ in Table 2 are based on the BART-large model, which obtained the best performance on both datasets. Consistent with results from passage-level ARA, the Neural Model achieved better performance over the Linguistic Model on both datasets in all metrics. Despite the relatively small amount of training data in the WSJ datasets, the Neural Model still offered competitive performance.

Hybrid Models. The previous best published result for the WSJ dataset 0.600, obtained with an

[7] The hyperparameters for learning rate, dropout and batch size are tuned on the dev set. We found best performance with learning rate at $1 \cdot e^{-5}$, dropout at 0.2, and set batch size as 32.

Dataset	Metric	BERT base	BART base	RoBERTa base	XLNet base	ELECTRA base	BART large	RoBERTa large	ELECTRA large
WSJ	Acc.	0.606	0.648	0.668	0.640	0.602	**0.679**	0.667	0.630
	F1	0.527	0.590	0.596	0.540	0.520	**0.611**	0.603	0.523
	Prec.	0.480	0.566	0.576	0.469	0.477	**0.601**	0.589	0.453
	Recall	0.606	0.648	0.668	0.640	0.602	**0.679**	0.667	0.630
	QWK	0.540	0.678	0.640	0.601	0.552	0.661	**0.677**	0.552
OSE	Acc.	0.547	0.571	0.569	0.562	0.555	**0.571**	0.570	0.566
	F1	0.532	0.555	0.554	0.543	0.533	**0.558**	0.555	0.549
	Prec.	0.549	0.570	0.566	0.554	0.552	0.565	**0.567**	0.566
	Recall	0.547	0.571	0.569	0.562	0.555	**0.571**	0.570	0.566
	QWK	0.500	0.537	0.537	0.535	0.512	**0.549**	0.541	0.532

Table 1: ARA performance of the Neural Model based on different transformers

Dataset	Metric	Linguistic Model	Neural Model	Hybrid Model		
				Concatenated	Hard Label	Soft Labels
WSJ	Acc.	0.618	0.679	0.629	**0.729**	0.724
	F1	0.549	0.611	0.590	0.707	**0.709**
	Prec.	0.519	0.601	0.585	0.713	**0.715**
	Recall	0.618	0.679	0.629	**0.729**	0.724
	QWK	0.616	0.661	0.676	0.767	**0.794**
OSE	Acc.	0.451	0.571	0.568	0.578	**0.581**
	F1	0.428	0.558	0.559	**0.565**	0.564
	Prec.	0.441	0.565	0.584	**0.593**	0.574
	Recall	0.451	0.571	0.568	0.578	**0.581**
	QWK	0.288	0.549	0.540	0.537	**0.560**

Table 2: ARA performance of the Linguistic Model, Neural Model (BART-large) and Hybrid Model

SVM model (Brunato et al., 2018). The Hybrid Model with Hard Label surpassed this result by almost 13% absolute to achieve state-of-the-art result, at 0.729 accuracy. The Soft Labels Model produced the second best performance, followed by the Neural Model. The Concatenated Model did not outperform the Neural Model, which may because long complex sequences and the size of dataset easily lead to overfit on the transformer-based models. The improvement of the Hard Label Model over the Neural Model[8] was statistically significant.

On the OSE dataset, the Soft Labels Model obtained the best performance in accuracy, though at a lower accuracy (0.581) than on the WSJ dataset. This likely reflects more fuzzy boundaries between the categories in the OSE corpus, where all sentences in the original texts were used. The Hard Label Model produced the second best performance as OSE dataset, followed by the Neural Model also. The Concatenated Model obtained worse performance than Neural Model also. The improvement of the Soft Label Model over the Neural Model[9] was statistically significant.

6 Conclusion

We have presented the first study on hybrid models on automatic readability assessment (ARA) at the sentence level. Our contribution is two-fold. First, we demonstrated that hybrid models outperform neural models, suggesting that linguistic features can capture salient properties that indicate sentence difficulty. Second, we compared three types of hybrid model, and showed that using the neural model's predictions as features in a traditional classifier yielded the best result, surpassing the previous best published result on the WSJ dataset by almost 13% absolute. These experimental results are expected to help inform future research on sentence-level ARA.

[8] At $p < 3.6 \cdot e^{-6}$ according to McNemar's Test.

[9] At $p < 1.4 \cdot e^{-4}$ according to McNemar's Test.

7 Limitation

Our experimental results should be interpreted with the following limitations in mind. First, our experiments involved relatively small datasets in English only. The performance of the model should also be evaluated on other languages and larger datasets. Second, the improvement observed in our best models depends on both the efficacy of the linguistic features and on the strength of the neural model itself. As neural models continue to improve and effective linguistic features are identified, the best methods for combining may also need to be updated.

Acknowledgement

This work was partly supported by the Language Fund from the Standing Committee on Language Education and Research (project EDB(LE)/PR/EL/203/14) and by the General Research Fund (project 11207320).

References

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing Relative Sentence Complexity using an Incremental CCG Parser. In *Proceedings of NAACL-HLT 2016*.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Kevyn Collins-Thompson. 2008. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. 2014. Assessing the Readability of Sentences: Which Corpora and Features? In *Proc. 9th Ninth Workshop on Innovative Use of NLP for Building Educational Applications*.

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: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, page 335–348. Springer.

Cristina Gârbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable Prediction of Text Complexity: The Missing Preliminaries for Text Simplification. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 611–618.

Adam Kilgarriff, Mils Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proc. EURALEX*.

Peter J. Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. In *Research Branch Report 8–75*. Chief of Naval Technical Training: Naval Air Station Memphis.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. 3rd International Conference for Learning Representations*, San Diego.

Bruce W. Lee, Yoo Sung Jang, and Jason Hyung-Jong 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*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Ho Hung Lim, Tianyuan Cai, John S. Y. Lee, and Meichun Liu. 2022. Robustness of hybrid models in cross-domain readability assessment. In *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, pages 62–67, Adelaide, Australia. Australasian Language Technology Association.

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*.

Dawei Lu, Xinying Qiu, and Yi Cai. 2020. Sentence-level readability assessment for l2 chinese learning. *CLSW 2019, LNAI*, 11831:381–392.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. International Conference on Learning Representations (ICLR)*.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 915–932.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, and O. Grisel. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *International Journal of Computational Linguistics and Applications*, 7(1):143–159.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and Machine Learning Approaches for Second Language Sentence-level Readability. In *Proc. 9th Workshop on Innovative Use of NLP for Building Educational Applications*.

Daniele Schicchi, Giovanni Pilato, and Giosué Lo Bosco. 2020. Deep neural attention-based model for the evaluation of italian sentences complexity. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 253–256. IEEE.

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1871–1881.

Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems*, 230:107408.

Yao-Ting Sung, Ju-Ling Chen, Ji-Her Cha, Hou-Chiang Tseng, Tao-Hsing Chang, and Kuo-En Chang. 2015. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*, 47:340–354.

Hou-Chiang Tseng, Hsueh-Chih Chen, Kuo-En Chang, Yao-Ting Sung, and Berlin Chen. 2019. An Innovative BERT-Based Readability Model. In *Lecture Notes in Computer Science, vol 11937*.

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.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proc. 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page 288–297.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic Assessment of Absolute Sentence Complexity. In *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proc. 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 141 – 153.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

A Appendix: Corpus statistics

WSJ		
Score	# sent	sent length
1	69	10.43
2	262	14.51
3	203	25.00
4	96	30.70
5	20	31.50
Total	650	20.27

Table 3: Size of the WSJ dataset and the average sentence length

OSE		
Version	# sent	sent length
Beginner	4,840	18.75
Intermediate	4,759	22.44
Advanced	4,632	25.90
Total	14,231	22.31

Table 4: Size of the OSE dataset and the average sentence length