Cambridge at SemEval-2021 Task 1: An Ensemble of Feature-Based and Neural Models for Lexical Complexity Prediction

Zheng Yuan   Gladys Tyen   David Strohmaier
ALTA Institute, University of Cambridge, United Kingdom
Department of Computer Science and Technology, University of Cambridge, United Kingdom
{firstname.lastname}@cl.cam.ac.uk

Abstract

This paper describes our submission to the SemEval-2021 shared task on Lexical Complexity Prediction. We approached it as a regression problem and present an ensemble combining four systems, one feature-based and three neural with fine-tuning, frequency pre-training and multi-task learning, achieving Pearson scores of 0.8264 and 0.7556 on the trial and test sets respectively (sub-task 1). We further present our analysis of the results and discuss our findings.

1 Introduction

Predicting which words are considered hard to understand for a given target population has many applications. For example, it can be used to identify texts appropriate for language learners or included in a pipeline for automatic text simplification for people with low literacy skills or reading disabilities (Xia et al., 2016; Shardlow, 2014; Gooding and Kochmar, 2019b). In this paper, we describe our submission to the SemEval-2021 shared task on Lexical Complexity Prediction (LCP) (sub-task 1), where participating teams are expected to predict the complexity score of single words in context (Shardlow et al., 2021). Compared to previous shared tasks on Complex Word Identification (CWI), which have primarily focused on binary classification as systems were expected to identify words as complex or not (Paetzold and Specia, 2016a; Yimam et al., 2018); a new multi-domain English dataset was used for the purpose, which was annotated using a 5-point Likert scale (Shardlow et al., 2020). We approached LCP as a regression problem and proposed a traditional feature-based model, as well as three neural models exploring fine-tuning, frequency pre-training and multi-task learning (MTL).

The remainder of this paper is organised as follows. Section 2 presents related work in the area. In Section 3, we describe our approach to the task and detail the four models included in our final ensemble system. In Section 4, we turn to the experiments, describing the data and evaluation metrics used, and presenting our results on the shared task trial set. Section 5 presents our official results on the shared task test set, and offers a discussion of the results and the performance of our submitted system. Finally, we conclude the paper and provide an overview of our findings in Section 6.

2 Related work

The SemEval-2016 shared task on CWI (Paetzold and Specia, 2016a) was framed as a binary classification problem, where complexity was defined as whether or not a word is difficult to understand for non-native English speakers. A set of 400 non-native speakers annotated the data in a binary fashion and a word was labelled as complex if it was annotated as complex by at least one annotator. The study performed by Zampieri et al. (2017) showed that most systems performed poorly due to the way the data was annotated. They also found out that words that were annotated as complex by the majority of human annotators tend to be easier for systems to identify, arguing that lexical complexity should be seen as a continuum on a spectrum rather than a binary value.

The second CWI shared task was organized as part of the BEA-2018 workshop (Yimam et al., 2018). It extended the previous one by introducing a new probabilistic classification sub-task where participants were asked to assign the probability that an annotator would find a word complex. The continuous complexity value for each word was calculated as the proportion of annotators that found a word complex. The results of the shared task showed that traditional feature engineering approaches (mostly based on length and

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 590–597
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

Linguistic feature	**Pearson**			**Importance**
	Train	**Trial**	**Test**	
1. Number of syllables in target word	**0.162**	**0.235**	**0.120**	0.019
2. Number of characters in target word	**0.099**	0.159	0.094	0.103
3. Mean number of syllables per word in given text	**0.057**	**0.224**	0.073	0.132
4. Mean number of characters per word in given text	**0.084**	**0.230**	0.089	0.119
5. Deviation from mean number of syllables	**0.079**	**-0.180**	-0.096	0.151
6. Deviation from mean number of characters	**-0.138**	-0.093	-0.062	0.180
7. Frequency of target word	**-0.240**	**-0.247**	**-0.183**	0.369
8. Frequency of target lemma	**-0.125**	**-0.256**	**-0.209**	0.714
9. Frequency of target dependency label	**-0.040**	-0.091	-0.054	0.059
10. Frequency of target POS	**0.051**	0.136	-0.010	0.128
11. Frequency of target word and dependency label	**-0.195**	-0.111	**-0.200**	0.089
12. Frequency of target word and POS	**-0.257**	**-0.201**	**-0.154**	0.126
13. Number of compounds in surrounding phrase	0.001	0.092	0.079	0.021
14. Number of modifiers in surrounding phrase	**0.073**	0.085	0.068	0.006
15. Number of dependencies linked to target word	**-0.073**	-0.044	-0.059	0.039

Table 1: Features used in the random forest regressor and the corresponding Pearson's correlation with complexity in the training (train), trial, and test data; as well as the corresponding mean permutation importance ($n = 50$). Bold Pearson values are significant ($p < 0.001$).

frequency features) performed better than neural network and word embedding approaches, including the winning system **Camb-2018** from Gooding and Kochmar (2018). However, this system was subsequently outperformed by a sequence labeller approach to CWI that incorporated word context (Gooding and Kochmar, 2019a). In both shared tasks, the top-performing systems demonstrated the strength of ensemble models (Paetzold and Specia, 2016b; Gooding and Kochmar, 2018).

3 Approach

3.1 Random forest regression

As a baseline, we began with training a simple random forest regressor based on 15 manually selected linguistic features. The regressor was trained with 100 trees, and we used mean absolute error (MAE) to measure the quality of each split. Most of our features were inspired by psycholinguistic studies and readability metrics. The full list of features can be found in Table 1.

Frequency Based on the psycholinguistic findings that the frequency of a word is strongly correlated with the speed at which it is processed (Preston, 1935; Monsell et al., 1989; Brysbaert et al., 2011), we introduced six features which are based on frequencies found in the Simple English Wikipedia (SimpleWiki).[1] We selected SimpleWiki for its standardised form, relatively low frequency of complex words, and coverage of a large number of topics. Two of our frequency-based features were calculated based on the frequency of words that match both the surface form and the syntactic role - this was done as a coarse form of word sense disambiguation, but also to capture syntactic complexity.

Syntax Psycholinguistic studies have shown that syntactic complexity is linked to processing speed (Ferreira, 1991) and working memory limitations (Norman et al., 1992), which may affect participants' perception of lexical complexity. In a similar vein, we added three syntactic features: the number of compounds and modifiers in the phrase containing the target word, and the number of child dependencies linked to the target word.

Readability We included syllable-based[2] and character-based metrics, which were inspired by traditional readability metrics such as the Flesch-Kincaid readability tests (Kincaid et al., 1975)

[1] The Simple English Wikipedia data can be accessed at https://simple.wikipedia.org/. For this shared task, we used a preprocessed version at https://github.com/LGDoor/Dump-of-Simple-English-Wiki.

[2] The number of syllables were estimated using the Syllables library: https://github.com/prosegrinder/python-syllables.

and the Coleman-Liau index (Coleman and Liau, 1975).

3.2 Fine-tuning BERT

Fine-tuning pre-trained language models via supervised learning has become the key to achieving state-of-the-art performance in various natural language processing (NLP) tasks. Our approach builds upon this, where we used BERT (Devlin et al., 2019) as the underlying language model and added a linear layer on top that allows for regression.

We treated it as a *sequence regression* problem and constructed the input by concatenating the target word w_t, the complexity of which was to be determined, and its context sentence:

$$[CLS]; w_t; [SEP]; w_1, ..., w_t, ...; [SEP] \quad (1)$$

We then fed the $[CLS]$ representation into the output layer for regression.

We used the L1-loss, which measures the MAE for the prediction, i.e.:

$$Loss = mean(\{l_1, \ldots, l_N\}); l_n = |x_n - y_n| \quad (2)$$

where x and y are respectively the output of the model and the target value. N is the batch size.

During training, the whole model was optimised in an end-to-end manner.

3.3 Frequency pre-training

We proposed an extension to the fine-tuning BERT system by introducing a pre-training step. We constructed a new pre-training set with 20,000 sentences extracted from SimpleWiki, filtering for whole sentences by detecting the presence of verbs, and removing sentences that are longer than 256 words, as this is the length of the longest sentence in the training data.

Frequency of each word and part-of-speech (POS) combination in SimpleWiki was counted and converted into a value between 0 and 1:

$$1 - \frac{\ln(f)}{\ln(h)} \quad (3)$$

where f is the original frequency value and h is the highest frequency found (excluding stop words). This conversion makes use of the Zipfian distribution observed in natural language (Zipf, 1935), allowing the model to be pre-trained on output values that match the range in the shared task dataset (see Section 4.1 for more details).

Data	Train	Trial	Test
Bible	2,574	143	283
Biomedical	2,576	135	289
Europarl	2,512	143	345
Total	7,662	421	917

Table 2: Number of single word instances in the training (train), trial and test subsets of the Bible, Biomedical and Europarl datasets.

We chose this particular frequency feature because it is the most strongly correlated one with the complexity values in the training data among the 15 features used in the random forest regressor (see Table 1 #12).

3.4 Neural multi-task learning

MTL allows models to use information from related tasks and learn from multiple objectives, which leads to performance improvement on individual tasks (Rei and Yannakoudakis, 2017; Yuan et al., 2019; Taslimipoor et al., 2020; Andersen et al., 2021). Instead of only predicting the complexity value of word in context, we extended the model to incorporate auxiliary objectives. We used a joint learning approach trained on in-domain data only and experimented with three related tasks to boost model performance:

- POS tagging
- Grammatical Relations (GR) prediction: We included as an auxiliary objective the prediction of the GR type of a dependent with its head.
- Genre classification: A classification task was introduced to predict the genre of the text.

Model weights were shared between the main and auxiliary training objectives. We used pre-trained DistilBERT (Sanh et al., 2020) for language representation as the basis for our neural network and added additional layers on top of the Transformer (Vaswani et al., 2017) architecture for fine-tuning.

The final layer for the LCP objective is a fully connected layer that performs regression. Different from the first two neural systems, we treated it as a *token regression* problem, where we only input the context sentence, and fed the vector representation of the target word w_t into the output layer for

Hyper-parameter	**BERT**	**BERT$_{\textbf{freq.}}$**	**MTL**
Language model	bert-base-uncased	bert-base-uncased	distilbert-base-uncased
Max. length	190	160	304
Batch Size	40	8	1
Epochs	5	7	4
Decay rate	0.01	0.01	0.01
Learning rate	5e-06	2e-05	1e-05
Schedule	linear	linear	linear
Warm up steps	80	90	7662

Table 3: Hyper-parameters used for experiments.

regression:

$$[CLS]; w_1, ..., w_t, ..., w_N; [SEP] \tag{4}$$

For those cases where the target word was split into multiple sub-tokens, we took the averaged representation.

Additionally, a new output layer was introduced to perform the auxiliary task. For the first two token-level auxiliary tasks (POS and GR), the token representations were fed into the output layer. The model only predicted labels for auxiliary objectives on the first token of a word, in an identical fashion to Devlin et al. (2019). For genre classification, we used the $[CLS]$ representation. The overall loss function is a weighted sum of the main LCP loss (measured as MAE) and the auxiliary loss (as cross-entropy):

$$Loss = \lambda Loss_{LCP} + (1 - \lambda) Loss_{aux} \tag{5}$$

4 Experiments

4.1 Dataset and evaluation

The data used in this shared task is an augmented version of CompLex (Shardlow et al., 2020), a multi-domain English dataset with sentences annotated using a 5-point Likert scale with 1 being very easy and 5 being very difficult. The final complexity labels were normalised in the range of $[0, 1]$. The dataset contains texts of three genres (Bible, Biomedical and Europarl) and both single words (sub-task 1) and multi-word expressions (sub-task 2). Since we focused on sub-task 1, we used only single word instances in our experiments. Corpus statistics are given in Table 2.

Systems were evaluated using Pearson correlation. We also report scores for the following metrics: Spearman correlation, MAE, mean squared error (MSE) and R-squared (R2).

4.2 Training details

We used spaCy[3] to preprocess the data and automatically generated lemma, POS and GR labels to be used in our experiments.

For the feature-based system, we used the random forest regressor in the *scikit-learn* library.[4] For the neural systems, we used pre-trained language models provided by *huggingface* (Wolf et al., 2020).[5] All neural systems were trained using the AdamW variant (Loshchilov and Hutter, 2019) of the Adam stochastic optimisation algorithm (Kingma and Ba, 2015). Detailed hyper-parameters are listed in Table 3. Each neural model was trained on one NVIDIA Tesla P100 GPU.

4.3 Individual system performance

Individual system performance on the trial set is reported in Table 4, where **RandomForest** refers to the feature-based random forest regression system, **BERT** refers to the fine-tuned BERT system, **BERT$_{\textbf{freq.}}$** refers to the fine-tuned BERT system with frequency pre-training, and **MTL$_{\textbf{X}}$** refers to the MTL system with the subscript '$_X$' representing the auxiliary task (POS, GR, or genre). We also report results from the winning system **Camb-2018** in the BEA-2018 CWI shared task, a feature-based, context-independent linear regression model.

We can see that our feature-based **RandomForest** system achieved comparable performance to the heavily feature-engineered **Camb-2018** system, despite using only 15 features. This may be due to the fact that linguistic features are often highly interdependent and capture very similar information.

We also notice that all our neural systems outperformed both feature-based systems by large margins (+0.1 Pearson). This contradicts the findings

[3] `https://spacy.io/`

[4] `https://scikit-learn.org/`

[5] `https://huggingface.co/transformers/`

System	Pearson	Spearman	MAE	MSE	R2
RandomForest	0.7043	0.6746	0.0751	0.0096	0.4934
BERT	0.7907	**0.7579**	0.0647	0.0072	0.6191
BERT$_{freq.}$	**0.8089**	0.7546	**0.0646**	**0.0068**	**0.6397**
MTL$_{POS}$	0.8000	0.7528	0.0662	0.0075	0.6052
MTL$_{GR}$	0.7936	0.7208	0.0654	0.0070	0.6290
MTL$_{genre}$	0.7982	0.7272	0.0656	0.0070	0.6300
Camb-2018	0.7079	0.6885	0.0746	0.0095	0.4957

Table 4: Performance of individual systems on the trial set (sub-task 1). The best results are marked in bold. **Camb-2018** is the winning system in the BEA-2018 CWI shared task.

Ensemble	Pearson	Spearman	MAE	MSE	R2
MTL$_{All}$	0.8129	0.7471	0.0634	0.0065	0.6542
MTL$_{All}$ + BERT + BERT$_{freq.}$	0.8228	0.7641	**0.0621**	**0.0063**	0.6684
MTL$_{All}$ + BERT + BERT$_{freq.}$ + RandomForest	**0.8264**	**0.7676**	0.0623	**0.0063**	**0.6688**

Table 5: Performance of ensemble systems on the trial set (sub-task 1). The best results are marked in bold.

from the BEA-2018 CWI shared task where traditional feature-based approaches performed better than neural network and word embedding approaches. This could possibly be explained by the use of pre-trained Transformer-based language models in our neural systems, as well as a different annotation scheme employed when constructing the CompLex dataset used for this shared task. Nevertheless, our findings appear to match the general trend in NLP where neural systems are overtaking feature-based models as the state of the art. All our neural systems produced comparable results: **BERT$_{freq.}$** yielded the best Pearson, MAE, MSE and R2 scores; while **BERT** yielded the best Spearman score.

4.4 Ensemble performance

We further averaged the outputs from individual systems to obtain an ensemble. Table 5 shows results for different system combinations. Overall, the best system consists of all our individual systems proposed in Section 3, including the feature-based **RandomForest** system; and achieved the best Pearson score of 0.8264, Spearman of 0.7676, MSE of 0.0063, and R2 of 0.6688. The ensemble of all neural systems yielded the best MAE of 0.0621.

5 Official results and discussion

Our submission to the LCP shared task (sub-task 1) is the result of our best system (in terms of Pearson), an ensemble of three neural and one feature-based systems **MTL$_{All}$ + BERT + BERT$_{freq.}$ + RandomForest**. The official results are reported in Table 6. Our final system achieved a Pearson score of 0.7556.

5.1 Per-genre performance

Using the Pearson correlation metric, the highest performance is obtained on the Biomedical data, followed by the Bible and Europarl data. On the MAE metric, however, the worst performance is found for the Biomedical data (see Table 6). We hypothesise that this might result from differences in the distribution of the lexical complexity scores. In particular, the scores for the Biomedical data appear to have a slightly larger interquartile range (see Appendix A, Figure A.1c).

5.2 Individual system contribution

To measure the contribution of each individual system to the overall performance, a number of ablation tests were performed, where one system was removed at a time. Results in Table 6 suggest that all neural systems have positive effects on the overall performance. Among them, **MTL$_{All}$** is the most effective one, whose absence is responsible for a 0.02 decrease in Pearson, followed by **BERT$_{freq.}$** and **BERT**. Interestingly, removing **RandomForest** yielded a better Pearson score of 0.7560, indicating that it is detrimental and brought performance down. This is inconsistent with our results on the trial set (see Table 5), where all systems contributed to the final system.

		Pearson	**Spearman**	**MAE**	**MSE**	**R2**
Official		0.7556	0.7105	0.0646	0.0070	0.5705
Genre	Bible	0.7475	0.7154	0.0662	0.0076	0.5493
	Biomedical	0.7763	0.7274	0.0745	0.0088	0.6025
	Europarl	0.7195	0.6699	0.0551	0.0049	0.5169
Ablated system	RandomForest	0.7560	0.7126	0.0647	0.0070	0.5685
	BERT	0.7523	0.7069	0.0650	0.0070	0.5655
	$\text{BERT}_{\text{freq.}}$	0.7515	0.7067	0.0652	0.0071	0.5633
	MTL_{All}	0.7371	0.6920	0.0669	0.0076	0.5335

Table 6: Official results of our submitted system on the test set (sub-task 1). Per-genre performance and ablation test results are included.

Analysis of RandomForest To understand why the feature-based regressor performed worse on the test data, we examined the correlation between each feature and the complexity scores in the training (train), trial, and test sets. Results in Table 1 show that several linguistic features (particularly #3, #4, and #10) are more strongly correlated with scores in the trial data compared to the test data, which may explain the discrepancy in our results. Although most features appear to have a small but significant correlation with complexity in the training data, many are not significant in the test data, likely due to the smaller sample size. This suggests that, while there may be some weak, overall correlation between these features and complexity, there is sufficient noise in the data that the relationship is negligible and unreliable when used to predict the complexity of a given word.

Additionally, we investigated the importance of each feature in the random forest regressor, as measured by the mean permutation importance (Breiman, 2001) - see Table 1. Our analysis reveals that the frequency of the target lemma (#8) is the most important one, followed by the frequency of the target word itself (#7). Both of these features are more strongly correlated with complexity in the trial data than either the training or test data, which also contributes to the inconsistency described above.

6 Conclusion

This paper presents our contribution to the SemEval-2021 shared task on LCP. We competed in sub-task 1 (single words) with an ensemble system combining three neural models and one feature-based model. Our analysis reveals that even though all three neural systems perform comparably, the MTL system contributed the most to the ensemble system. Adding the feature-based model improved the performance on the trial data, but brought performance down on the test data. In addition to the mismatch between the trial and test data, the noise in the data further contributed to this inconsistency. The comparatively lower performance of the feature-based system is especially interesting in light that such systems were competitive for CWI until relatively recently (Gooding and Kochmar, 2018). When looking at different genres, our submitted system yielded the highest performance in Pearson, but worst performance in MAE in Biomedical domain, compared to the other genres. We hypothesise that this is due to differences in data distribution between genres.

Acknowledgments

We thank Sian Gooding and Ekaterina Kochmar for support and advice. This paper reports on research supported by Cambridge Assessment, University of Cambridge. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council. We acknowledge NVIDIA for an Academic Hardware Grant.

References

Øistein E. Andersen, Rebecca Watson, Zheng Yuan, and Kevin Yet Fong Cheung. 2021. Benefits of alternative evaluation methods for automated essay scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*. International Educational Data Mining Society.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental Psychology*, 58(5):412–424.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

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.

Fernanda Ferreira. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30(2):210–233.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019a. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019b. Recursive context-aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

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, Defense Technical Information Center.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR 2019)*.

Stephen Monsell, Michael C Doyle, and Patrick N Haggard. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1):43–71.

Suzanne Norman, Susan Kemper, and Donna Kynette. 1992. Adults' reading comprehension: Effects of syntactic complexity and working memory. *Journal of Gerontology*, 47(4):P258–P265.

Gustavo Paetzold and Lucia Specia. 2016a. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016b. SV000gg at SemEval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974, San Diego, California. Association for Computational Linguistics.

Katherine A. Preston. 1935. The speed of word perception and its relation to reading ability. *The Journal of General Psychology*, 13(1):199–203.

Marek Rei and Helen Yannakoudakis. 2017. Auxiliary objectives for neural error detection models. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 33–43, Copenhagen, Denmark. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*, Vancouver BC, Canada.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, San Diego, CA. Association for Computational Linguistics.

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*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

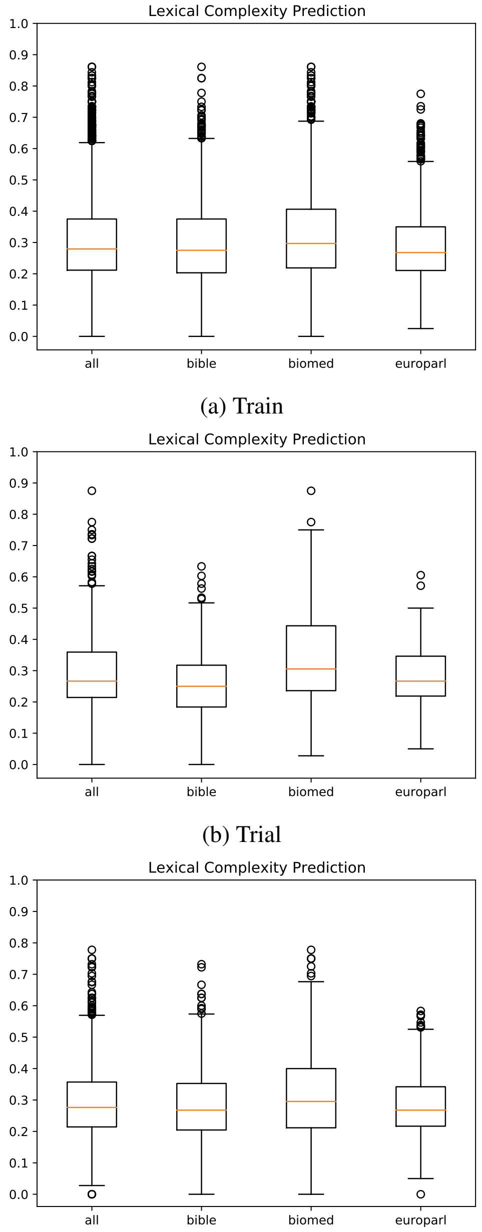
Zheng Yuan, Felix Stahlberg, Marek Rei, Bill Byrne, and Helen Yannakoudakis. 2019. Neural and FST-based approaches to grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 228–239, Florence, Italy. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

George Kingsley Zipf. 1935. *The Psycho-Biology Of Language*. Routledge.

A Data distribution

Figure A.1 presents the box plots of complexity scores in the training (train), trial and test subsets of the Bible, Biomedical and Europarl datasets.



(a) Train

(b) Trial

(c) Test

Figure A.1: Box plots of complexity scores for data by genre.