CSECU-DSG at SemEval-2021 Task 1: Fusion of Transformer Models for Lexical Complexity Prediction

Abdul Aziz, MD. Akram Hossain, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

Lexical complexity prediction (LCP) conveys the anticipation of the complexity level of a token or a set of tokens in a sentence. It plays a vital role in the improvement of various NLP tasks including lexical simplification, translations, and text generation. However, multiple-meaning of a word in multiple circumstances, grammatical complex structure, and the mutual dependency of words in a sentence make it difficult to estimate the lexical complexity. To address these challenges, SemEval-2021 Task 1 introduced a shared task focusing on LCP and this paper presents our participation in this task. We proposed a transformer-based approach with sentence pair regression. We employed two fine-tuned transformer models including BERT and RoBERTa to train our model and fuse their predicted score to the complexity estimation. Experimental results demonstrate that our proposed method achieved competitive performance compared to the participants' systems.

1 Introduction

Lexical complexity prediction (LCP) has become an important task in this globalization age, especially for second language learners (Przybyła and Shardlow, 2020). LCP is a little bit expansion of complex word identification (CWI) task (Paetzold and Specia, 2016; Štajner et al., 2018), where CWI is a binary classification of a word that is complex or not and LCP is finding the complexity level of a word in continuous labelling in a sentence (Shardlow et al., 2020). LCP plays a vital role in many NLP applications such as lexical simplification (Paetzold, 2016; Paetzold and Specia, 2017; Qiang et al., 2020), text generation, and machine translation (Wang et al., 2016). Besides, it helps those people who are suffering from Dyslexia (Rello et al., 2013a), Aphasia (Rello et al., 2013b), and those with low literacy levels (Aluisio and Gasperin, 2010).

LCP is a very challenging task (Zampieri et al., 2017), especially because the non-identical target audiences will have distinct needs. For example, speakers of one language usually less familiar with different subsets of the vocabulary of a second language. Besides, the grammatical shape of a sentence and the ambiguous meaning of a word in different places make this task more challenging and important to explore. A single word may portray different lexical complexity because of its non-identical usage, position, tense form, and redundancy in different sentences or in the same sentence. To estimate multi-word complexity, we need to consider the dependency between tokens.

Sentence	Token	Complexity
Sub-task 1		
His head is like the purest gold.	gold	0.210
Sub-task 2		
They shall eat it with bitter herbs.	bitter herbs	0.25

Table 1: Example of sub-task 1 and sub-task 2.

To address the challenges of lexical complexity prediction of words in sentences, (Shardlow et al., 2021a) proposed a shared task at SemEval-2021 Task 1. The task is divided into two subtasks. In sub-task 1, a system needs to determine the complexity level of a word in the sentence, whereas in sub-task 2, a system needs to determine the overall complexity level of multiple words in the sentence. To explain the definition of both sub-tasks, we articulate a few examples in Table 1.

**The first two authors have equal contributions.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 627–631
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

Taking part in the LCP shared task of SemEval-2021, we exploit the pairwise contextual information of sentence and token. In this regard, we proposed a combined transformer based framework with sentence pair regression. We make a pairwise learning framework with the sentence-token pair to train the two state-of-the-art transformers model including BERT and RoBERTa.

We organize the rest of the paper as follows: Section 2 presents the details of our proposed framework. Whereas in Section 3, we present our experimental settings and analyze the performance of our model against the various settings and related methods. Finally, we conclude our paper in Section 4 with some future directions.

2 Proposed Lexical Complexity Prediction Framework

In this section, we describe our proposed lexical complexity prediction framework. Our goal is to predict the complexity score of a token or a set of tokens in the given sentence. We depict the overview of our framework in Figure 1.

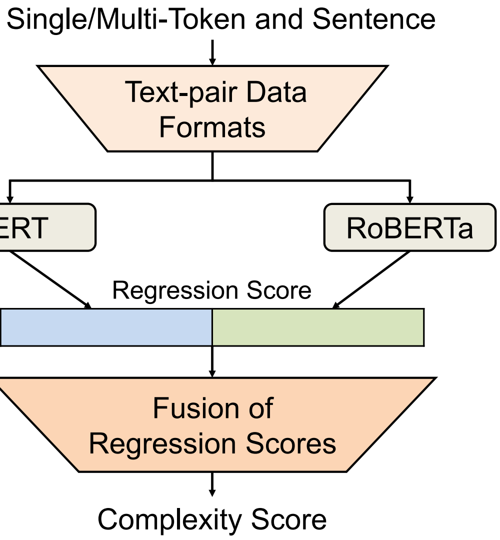

Figure 1: Overview of our proposed framework.

In our framework, we use a sentence pair regression concept in transformer models to perform lexical complexity prediction where input sentence and target word pairs are packed together into a single sequence. After performing sentence-token pair regression through BERT and RoBERTa models, we estimate each model's regression score. Subsequently, we fuse these models' predictions by taking the mean of these scores to determine the final complexity score.

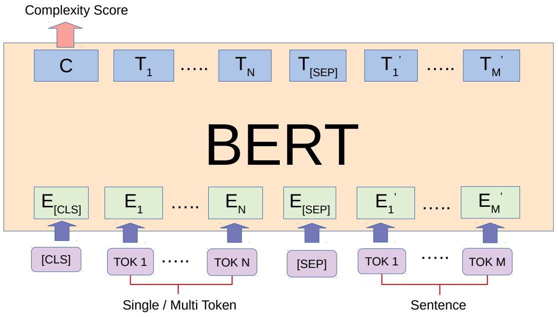

Figure 2: Pairwise learning using BERT model.

2.1 Fine-tuned Transformer Models

We fine-tune the transformer models to perform sentence pair regression for LCP through BERT and RoBERTa. We describe the details in the subsequent sections.

2.1.1 Input Representation

We train with the sentence-word pair for better understanding their contextual relation which in turn helps to estimate the complexity of the target word in the sentence. It is important for an LCP system to predict both single and multi-words complexity. We exploit Huggingface transformers library (Wolf et al., 2020) with pairwise training where input target words and sentence make pair as a single sequence and detached with the [SEP] token. We utilize two pre-trained transformer models including RoBERTa and BERT. For LCP tasks training, each model's first token is the special [CLS] token at the beginning of every sequence which is also responsible for the final layer regression score of each model. For each sequence, we separate every pair with [SEP] token (as presented in Figure 2) where the target words belong to text_a and sentence belongs to text_b. We fine-tune the architecture with the pre-trained BERT and RoBERTa models to estimate the complexity score.

2.1.2 BERT

BERT (Devlin et al., 2019) stands for bidirectional encoder representations from transformers, is a new method of pre-training sentence representations which achieves state-of-the-art results on many NLP tasks including question-answering, text classification, and sentence-pair regression. We take advantage of the bert fast tokenizer and bert-base-uncased model for sentence-pair regression where target words and sentence make pair as a single sequence.

2.1.3 RoBERTa

RoBERTa (Liu et al., 2019) is an extension to the original BERT model which is named as a robustly optimized BERT pre-training approach. It focuses on the key hyper-parameters choices and removing the next sentence prediction (NSP) objective. Besides, it is training with much larger mini-batches and learning rates. We exploit the roberta fast tokenizer and roberta-base model for sentence-pair regression to get the complexity score where target word and sentence are trained as pairwise training.

2.2 Fusion of Transformer Models

To ameliorate the performance of individual models, we fuse the predicted complexity score of two models to generate a unified score. We use the arithmetic mean to average both model's regression scores to determine the final complexity score. The estimation is computed as follows:

$$C_i = \frac{x_i + y_i}{2}$$

where x_i and y_i correspond to the BERT and RoBERTa regression score, respectively.

3 Experiment and Evaluation

3.1 Dataset Description

The organizers of the lexical complexity prediction (LCP) task 1 at SemEval-2021 (Shardlow et al., 2021a) provided a multi-domain English benchmark dataset (Shardlow et al., 2020, 2021b) to evaluate the performance of the participants' systems. The dataset was collected from three different corpuses including the Bible, europarl, and biomedical. The proposed task is divided into two subtasks, sub-task 1 focused on single word instances whereas sub-task 2 focused on multi-word instances. The training set for sub-task 1 contains 7662 instances where 2574 instances from Bible, 2576 instances from biomed, and 2512 instances from europarl. The training set of sub-task 2 comprises 1517 instances (505 Bible, 514 biomed, and 498 europarl). The validation set consists of 99 multi-word expressions (29 Bible, 33 biomed and 37 europarl) and 421 single word instances (143 Bible, 135 biomed and 143 europarl). The organizer provided 917 single word instances (283 Bible, 289 biomed, and 345 europarl) for sub-task 1 and 184 multi-word instances (66 Bible, 53 biomed, and 65 europarl) for sub-task 2 as a test set.

3.2 Experimental Settings

We now describe the set of parameters that we have used to design our proposed lexical complexity prediction model. In our CSECU-DSG system, we utilize two state-of-the-art Huggingface transformer models with fine-tuning, including BERT and RoBERTa. We use simpletransformers API (Rajapakse, 2019) to implement our system. We train our system with the provided training data. We trained BERT and RoBERTa model using 5 epochs and set the learning rate of 2.99e-5, save_steps = 767, and evaluate_during_training_steps = 40. We used the CUDA-enabled GPU and set the manual_seed = 4 to generate the reproducible results. Default settings were used for the other parameters.

3.3 Evaluation Measures

To evaluate the performance of participants' lexical complexity prediction systems, SemEval-2021 task 1 organizers used different strategies and metrics for sub-task 1 and sub-task 2 (Shardlow et al., 2020). For both sub-task, standard evaluation metrics including Pearson correlation (R), Spearman correlation (Rho), mean absolute error (MAE), mean squared error (MSE), and R-squared (R^2) were applied to estimate the performance of a system. However, Pearson correlation (R) is considered as the primary evaluation measure for both subtasks of this task.

3.4 Results and Analysis

The comparative results of our proposed CSECU-DSG system along with top-5 performing systems (Shardlow et al., 2021a) in sub-task 1 and sub-task 2 are presented in Table 2 and Table 3, respectively. Following the benchmark of SemEval-2021 task 1, participants' systems are ranked based on the primary evaluation metric Pearson correlation (R) score.

At first, we presented the performance of our proposed method. We also presented the performance of top-5 ranked participating systems and LCP baselines. Here, we see that our proposed method obtained competitive performance against the other top-performing systems. In comparison to the other participants' methods, we have seen that our system demonstrated a similar kind of performance on both sub-task. This deduces the applicability and generalizability of our system for the complexity estimation of both the single and multi-words.

Team (Rank)	Pearson	Spearmen	MAE	MSE	R^2
CSECU-DSG (9th)	0.7716	0.7326	0.0632	0.0066	0.5909
Top performing team based on Pearson correlation score					
JUST BLUE (1st)	0.7886	0.7369	0.0609	0.0062	0.6172
DeepBlueAI (2nd)	0.7882	0.7425	0.0610	0.0061	0.6210
Alejandro Mosquera (3rd)	0.7790	0.7355	0.0619	0.0064	0.6062
Andi (4th)	0.7782	0.7287	0.0637	0.0064	0.6036
CS-UM6P (5th)	0.7779	0.7366	0.0803	0.0100	0.3813

Table 2: Comparative results with other selected participants (Sub-task 1).

Team (Rank)	Pearson	Spearmen	MAE	MSE	R^2
CSECU-DSG (12th)	0.8311	0.8153	0.0678	0.0077	0.6825
Top performing team based on Pearson correlation score					
DeepBlueAI (1st)	0.8612	0.8526	0.0616	0.0063	0.7389
rg_pa (2nd)	0.8575	0.8529	0.0672	0.0072	0.7035
xiang_wen_tian (3rd)	0.8571	0.8548	0.0675	0.0072	0.7012
andi_gpu (4th)	0.8543	0.8448	0.0664	0.0071	0.7055
ren_wo_xing (5th)	0.8541	0.8473	0.0677	0.0073	0.6967

Table 3: Comparative results with other selected participants (Sub-task 2).

3.5 Discussion

In order to estimate the effect of each component of our CSECU-DSG model, we estimated the performance of the individual model. The summarized experimental results for sub-task 1 and sub-task 2 are presented in Table 4.

From the results, it can be observed that RoBERTa based model performed better compared to the BERT model when considering individual model's performances. However, combining two models regression scores by using mean increased Pearson correlation score by more than 1% on both subtasks. This deduced the importance of our model fusion.

All three models performed better for multi-words complexity estimation compared to the single word complexity. We have seen a similar kind of trend in other models' performances reported in Table 2, and Table 3. This demonstrated that estimating the single word complexity is more challenging compared to the multi-words expression. This is because a multi-word expression contains more words, therefore, contains more contextual information that helps the model for complexity estimation compared to the single word.

Method	Single Word	Multi Word
	Pearson	Pearson
CSECU-DSG	0.7716	0.8311
Performance of Individual Model		
−BERT	0.7514	0.8077
−RoBERTa	0.7634	0.8211

Table 4: Performance analysis of individual model.

4 Conclusion and Future Directions

In this paper, we presented our approach to the lexical complexity prediction task. We tackled the problem by performing sentence pair regression using two SOTA transformer models including BERT and RoBERTa in a unified architecture. By using pairwise learning, we exploited the contextual relation between sentence-word pairs to estimate the complexity score. Our method achieved competitive scores compared to other participants.

In the future, we have a plan to incorporate various handcrafted features with state-of-the-art neural methods to distill the relationship of sentence-word pairs for complexity estimation.

References

Sandra Aluisio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the PorSimples project for simplification of Portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53.

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 (NAACL:HLT)*, pages 4171–4186.

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 e-prints*, pages arXiv–1907.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

Gustavo H Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield.

Piotr Przybyła and Matthew Shardlow. 2020. Multi-Word Lexical Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1435–1446.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34,05, pages 8649–8656.

T. C. Rajapakse. 2019. Simple Transformers. https://github.com/ThilinaRajapakse/simpletransformers.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013a. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219. Springer.

Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013b. One half or 50%? An eye-tracking study of number representation readability. In *IFIP Conference on Human-Computer Interaction*, pages 229–245. Springer.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex: A New Corpus for Lexical Complexity Predicition from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021a. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021)*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2021b. Predicting Lexical Complexity in English Texts. *arXiv preprint arXiv:2102.08773*.

Sanja Štajner, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Anaïs Tack, Seid Muhie Yimam, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016. Text simplification using neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30,1.

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. 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.

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)*.