JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models

Tuqa Bani Yaseen Qusai Ismail Sarah Al-Omari Eslam Al-Sobh Malak Abdullah
Jordan University of Science and Technology
Irbid Jordan
tmbaniyaseen19, qfismail20, ssalomari18, esalsobh19 @ cit.just.edu.jo
mabdullah@just.edu.jo

Abstract

Predicting the complexity level of a word or a phrase is considered a challenging task. It is even recognized as a crucial step in numerous NLP applications, such as text rearrangements and text simplification. Early research treated the task as a binary classification task, where the systems anticipated the existence of a word's complexity (complex versus uncomplicated). Other studies had been designed to assess the level of word complexity using regression models or multi-labeling classification models. Deep learning models show a significant improvement over machine learning models with the rise of transfer learning and pre-trained language models. This paper presents our approach that won the first rank in the SemEval-task1 (sub stask1). We have calculated the degree of word complexity from 0-1 within a text. We have been ranked first place in the competition using the pre-trained language models BERT and RoBERTa, with a Pearson correlation score of 0.788.

keywords: Neuro-linguistic programming (NLP), Lexical Complexity Prediction(LCP), Deep Learning, RoBERTa, BERT.

1 Introduction

Lexical complexity plays a significant role in the readability level and comprehension. The precise anticipation of lexical complexity can help systems direct the user to an acceptable simple text accurately or modify the text to be more fluid (Brothers and Traxler, 2016). Predicting the complexity of words is a subjective and challenging problem, while it is conjectural, too. Yet, mapping words into their complexity is an essential task to understand natural language. Numerous components can influence the prediction of lexical complexity. Several approaches were proposed to solve or mitigate this type of study using Machine and Deep learning methods (Sengupta et al., 2020; Gooding and Kochmar, 2019; Bahja, 2020).

This paper describes the JUST-BLUE team's model that participated in the SemEval 2021-task1, Lexical Complexity Prediction (LCP) (Shardlow et al., 2021). The task provides participants with an augmented version of CompLex, a multi-domain English dataset with sentences annotated using a 5-point Likert scale (1-5) (from very easy to very difficult) (Shardlow et al., 2020). The task is to predict the complexity value of words in context. It is worth mentioning that our model, JUST-BLUE, has been ranked first in this task. We have used the pre-trained language models, BERT and RoBERTa Which have proven their effectiveness in this area (Liu et al., 2019), along with the ensembling method (weighted averaging) to achieve the highest Pearson correlation score of 0.788.

The rest of this paper is organized as follows: Section 2 sheds light on related work. Section 3 describes the methodology proposed in this research. Section 4 discusses the experimentation setup and evaluation results. Whereas Section 5 concludes this research.

2 Related work

One of the most prominent challenges in the current era is the prediction of lexical complexity. Prediction of the word complexity in machine learning can be binary; the word is complex or not complex. It also can be a non-binary prediction, as a probabilistic prediction with the measurement of complexity within a particular scale (0.6 the probability that the word is complex). SemEval 2016 introduced the first shared task of predicting word complexity with a mission limited to the word orders being complex or non-complex (binary prediction) (Paetzold and Specia, 2016). Decision Tree classifiers achieved the best results (Zampieri

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 661–666
Bangkok, Thailand (online), August 5–6, 2021.

et al., 2017). It has been noted that word length is a good indication of word complexity (De Hertog and Tack, 2018).

The authors in (Shardlow, 2013) discussed the importance of frequency and length of words. They used the Keras deep learning library to predict whether an English or Spanish word is complex or not. They used character embedding, word length, frequency count, word embedding, and psychological measures as features to predict complex words and achieved 0.872 as F1-score. The authors in (Yimam et al., 2018) worked on various languages, such as English, Spanish, French and German. They worked on two different methods for predicting complex words. The first method is to find if the word orders are either complex or simple. The second is to find the probability that the word is complex. The complex levels depended on the average of the annotators' answers. For example, if the number of annotators who expected the word to be complex is 6 out of 10, then the probability is 0.6. A claim stated that this annotating method is considered impractical since the probability of 0.5 cannot be considered complex or not complex. So the authors in (Shardlow et al., 2020) suggested a Likert scale with 5-point. The authors asserted that this method is more accurate scale instead of calling the word complex and non-complex. We can divide the word into being very easy, easy, neutral, difficult, and very difficult. This scale is beneficial to our work.

The deep learning pre-trained language models, BERT and RoBERTa, are considered state-of-the-art for NLP. Teams in the previous shared tasks of SemEval 2020 had used these models to obtain the best results for different NLP tasks (Al-Khdour et al., 2020; Shatnawi et al., 2020; Jurkiewicz et al., 2020). Our approach experimented with these models using different hyperparameters and weighted averaging methods that lead to the best result in the competition for predicting lexical complexity.

3 Methodology

This section describes our approach methodology and goes as follows: First, we describe the task's dataset. Then, the preprocessing step. Finally, we describe the JUST-BLUE approach to predict the word's complexity.

3.1 Data

The SemEval-task 1 competition has provided the contestants with three files (trial, train, and test data). The files contain several columns as follows:

- id: the identification number for each entry.
- corpus: the sources from which the words were being collected. It was extracted from three sources: the bible, biomedical, and The European Parliament.
- sentence: the set of words for which complexity needed to be measured.
- token: the single word in which complexity needed to be measured.
- complexity: the degree of complexity of the word, ranging from 0 to 1.

3.2 Pre-Processing Step

First, we cleaned the data and removed all single and double quotations manually. This step helped to separate some of the merged rows. Next, we deleted any row where columns contain the NaN value because it will not be effective in the training process.

3.3 JUST-BLUE Architecture

We have used the pre-trained language models, BERT and RoBERTa models. We have imported the BERT model using BERT-sklearn library as it includes SciBERT and BioBERT models for the scientific and biomedical fields. We also have used simple transformers; classification libraries to import the RoBERTa model. As we mentioned earlier, the goal of the task is to determine the complexity of the word. Knowing that the word's complexity changes slightly based on the complexity of the sentence, we have used both the token (word) and the sentence to predict the word's complexity. We have fed BERT and RoBERTa models with the 'token,' and the 'complexity' label to be trained. We have also inserted 'sentence' and 'complexity' columns to both models for training as a second strategy. The results have been combined using an ensembling voting method, Weighted Averaging. Our experiments show that the 80:20 ratio for weights can achieve the best results. The highest voting rate is for the "token" model (model 1) since we need to calculate the degree of complexity for a single word. On the other hand, the complexity of a

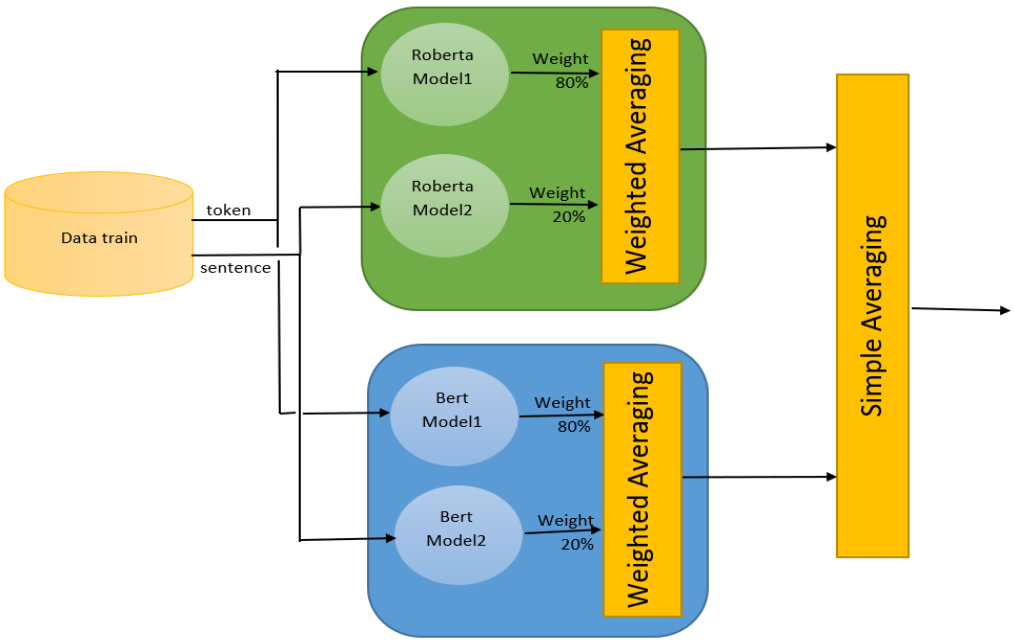

Figure 1: JUST-BLUE workflow

word is affected by the complexity of the sentence in which it is included. So, we gave a 20

The Simple Averaging method has been used as the ensembling technique to merge BERT and RoBERTa's models' results. Figure 1 illustrates the methodology used.

For more clarification, suppose we have the word 'sea' for which we want to calculate the complexity. The 'sea' word exists in this sentence "and they entered into the boat, and were going over the sea to Capernaum." First, we feed the word sea to model1 using RoBERTa. We also feed the sentence that contains the word sea to RoBERTa model2. Then, we combine the two results obtained using Weighted Averaging. Suppose that the RoBERTa model1 result is 0.01 (the word sea has a 0.01 complexity degree) and RoBERTa model2 is 0.13 (the sentence has a 0.13 complexity degree). The resulted RoBERTa models is 0.01x80% + 0.13x20%, which is equal to 0.034. We repeat these steps for BERT's models. If the BERT model has a result of 0.052, then the final step is to calculate the average of the RoBERTa and BERT model. The complexity is (0.034 + 0.052)/2, equal to 0.043, as shown in Figure 2.

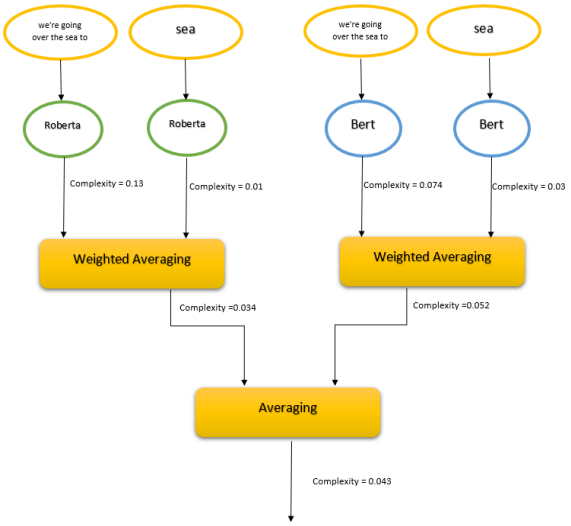

Figure 2: Example Description

4 Results and Discussion

We used Python version 3.6 on the Colab environment to execute our codes. We have experimented with several models to determine which models are suitable for this task. We have experimented with BERT and RoBERTa pre-trained models. We also examined SVM and Random Forest machine learning models. Table 1 shows the results we have obtained throughout our experiments.

The challenging step was to find the best weights for the models that used tokens (single words) and sentences to get the best result (Table2). As we mentioned earlier, some words have a different complexity degree, depending on their location in the sentences. Therefore, it was necessary to insert

Table 1: Results of different models

Models	score
SVM	0.3472
Random Forest	0.4503
BERT	0.8199
RoBERTa	0.8268
BERT and RoBERTa	0.8190

Table 2: Different weights (tokens and sentences)

model		score
Token %	Sentence %	
90	10	0.8258
80	20	0.8268
70	30	0.8252

each of the words and sentences for the training to verify the best weight. Table 2 shows the best weight, which is 80% for words and 20% for sentences.

The next step was to explore BERT and RoBERTa's best hyperparameters, such as learning rate, batch size, epochs, and max sequence length. Table 3 shows the description of these hyperparameters, and Table 4 shows example results of fine-tuning JUST-BLUE hyperparameters.

Finally, we thought of determining the effects of the base size and large size models of BERT and RoBERTa on the accuracy. It is shown by our experiments that the large sizes decreased the accuracy.

In the testing phase, we noticed that the words (tokens) in the file were new. Therefore, we decided to limit the number of arguments to avoid overfitting. We just changed "num-train-epochs "=3 in BERT and RoBERTa's model, but the other arguments had the default values. We have used three different models. The first was the BERT model, the second was the RoBERTa model, and the third was BERT and RoBERTa together as described in the Methodology Section. Table 5 shows the results we received from the different models we used.

JUST-BLUE approach achieved the best result using RoBERTa and BERT's models with a Pearson correlation of 0.788 scores. We have also achieved the least Mean Absolute Error(MAE) with 0.0609. Our model is ranked first the LCP-sub task1 of a single word. The Spearman's Rho (Rho) and R-squared (R2) scores are 0.7369 and 0.6172, respectively. The number of teams in the shared task Lexical Complexity Prediction (LCP) was 54 teams. This shared task is considered a high level of CWI 2016 and CWI 2018 with a larger number of words from various sources.

5 Conclusion

Predicting the complexity of words is one of the most prominent tasks that the NLP research community strives to solve. It is worth noting that in 2016 and 2018, two tasks were issued to determine whether the word was complex or not. SemEval 2021 introduced task 1, Lexical Complexity Prediction (LCP) that aims to predict the word's complexity from 0 to 1. This paper described the top-ranked team's model, JUST-BLUE. The JUST-BLUE model obtained the highest Pearson Correlation score of 0.788 using the pre-trained language models BERT and RoBERTa. Our strategy depends on the ensembling methods, Simple and Weighted Averaging.

References

Nour Al-Khdour, Mutaz Bni Younes, Malak Abdullah, and AL-Smadi Mohammad. 2020. Justmasters at SemEval-2020 task 3: Multilingual deep learning model to predict the effect of context in word similarity. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 292–300.

Mohammed Bahja. 2020. Natural language processing applications in business. In *E-Business*. IntechOpen.

Trevor Brothers and Matthew J Traxler. 2016. Anticipating syntax during reading: Evidence from the boundary change paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(12):1894.

Dirk De Hertog and Anaïs Tack. 2018. Deep learning architecture for complexword identification. In *Thirteenth Workshop of Innovative Use of NLP for Building Educational Applications*, pages 328–334. Association for Computational Linguistics (ACL); New Orleans, Louisiana.

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

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. Applicaai at SemEval-2020 task 11: On roberta-crf, span cls and whether self-training helps them. *arXiv preprint arXiv:2005.07934*.

Table 3: Description of some argument

Argument	Description
learning rate	The learning rate for training
number of epochs	One forward pass and one backward pass of training examples
max sequence length	Maximum sequence length the model will support
train batch size	Determines the number of samples in each mini batch
early stopping metric	Stop training when early stopping metric doesn't improve
early stopping delta	Counts as a better checkpoint
evaluate during training steps	Performs evaluation at every specified number of steps

Table 4: Hyperparameter Fine-Tuning for JUST-BLUE model in the training phase

#	*learning rate*	*number of epochs*	*max sequence length*	*train batch size*	*early stopping metric*	*early stopping delta*	*evaluate during training steps*	*score*
1	4e-5	1	128	8	mcc	0	2000	0.796245
2	5e-5	1	64	16	eval_loss	0.01	1000	0.799285
3	4e-5	2	128	16	mcc	0.01	2000	0.80589
4	4e-5	3	128	8	eval_loss	0	2000	0.819013
5	4e-5	3	64	8	mcc	0.01	2000	0.827414
6	5e-5	3	32	16	eval_loss	0.02	1000	0.784574
7	4e-5	5	128	16	mcc	0	2000	0.815554
8	3e-5	5	64	8	eval_loss	0	500	00.818949
9	4e-5	4	128	8	mcc	0.01	1000	0.827172
10	3e-5	4	64	16	eval_loss	0.02	2000	0.82385

Table 5: Results of different models in the testing phase

Model	**Score**
RoBERTa	0.7772
BERT	0.7556
JUST-BLUE (RoBERTa and BERT)	0.7886

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*.

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)*, pages 560–569.

Saptarshi Sengupta, Sanchita Basak, Pallabi Saikia, Sayak Paul, Vasilios Tsalavoutis, Frederick Atiah, Vadlamani Ravi, and Alan Peters. 2020. A review of deep learning with special emphasis on architectures, applications and recent trends. *Knowledge-Based Systems*, 194:105596.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. Complex—a new corpus for lexical complexity predicition from likert scale data. *arXiv preprint arXiv:2003.07008*.

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

Fara Shatnawi, Malak Abdullah, and Mahmoud Hammad. 2020. Mlengineer at SemEval-2020 task 7: BERT-flair based humor detection model (bfhumor). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1041–1048.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. *arXiv preprint arXiv:1710.04989*.