IAPUCP at SemEval-2021 Task 1: Stacking Fine-Tuned Transformers is Almost All You Need for Lexical Complexity Prediction

Kervy Rivas Rojas
Research Group on Artificial Intelligence
Pontificia Universidad Católica del Perú
k.rivas@pucp.edu.pe

Fernando Alva-Manchego
Department of Computer Science
University of Sheffield
f.alva@sheffield.ac.uk

Abstract

This paper describes our submission to SemEval-2021 Task 1: predicting the complexity score for single words. Our model leverages standard morphosyntactic and frequency-based features that proved helpful for Complex Word Identification (a related task), and combines them with predictions made by Transformer-based pre-trained models that were fine-tuned on the Shared Task data. Our submission system stacks all previous models with a LightGBM at the top. One novelty of our approach is the use of multi-task learning for fine-tuning a pre-trained model for both Lexical Complexity Prediction and Word Sense Disambiguation. Our analysis shows that all independent models achieve a good performance in the task, but that stacking them obtains a Pearson correlation of 0.7704, merely 0.018 points behind the winning submission.

1 Introduction

Complex Word Identification (CWI) consists of determining which words or multi-word expressions (MWE) in a text could be difficult to understand by certain readers. This is one of the first steps in the typical Lexical Simplification pipeline (Shardlow, 2014). CWI has traditionally been treated as either a binary (Paetzold and Specia, 2016) or regression (Štajner et al., 2018) task. For the latter, the complexity of a word/MWE was computed as a percentage of binary complexity ratings. Recently, Shardlow et al. (2020) proposed to move away from the binary definition of CWI, and instead collected complexity ratings using Likert scales. This allows re-defining the task as Lexical Complexity Prediction (LCP). Leveraging this new collected data, the First LCP Shared Task was organised in SemEval-2021 (Shardlow et al., 2021).

Our team participated in Sub-task 1: predicting the complexity score of single words. Basically, given a sentence and a target word in it, the goal is to predict the complexity score of the target. One particular challenge is that the same target can have different complexity scores depending on the sentence it appears in. Therefore, our proposed approach takes the context of the target into consideration in two ways. First, we use contextualised word representations from pre-trained Transformed-based models, such as RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019). In particular, we use the LCP data to fine-tune two RoBERTa models and one XLNNet model that receive as input the target and a context window of 1, and a RoBERTa model whose inputs are the target and a context window of 2. Second, we hypothesise that different contexts could evoke different senses of the target word. As such, we exploit data for Word Sense Disambiguation (WSD) through multi-task learning. In particular, we fine-tune a BERT (Devlin et al., 2019) model with two tasks: LCP and WSD, using the Unified Evaluation Framework (Raganato et al., 2017) for the latter. The predictions from all these models are combined with several morphosyntactic and corpus-based features, and used to train a Gradient Boosting Decision Tree with LightGBM (Ke et al., 2017).

On the test set of the Shared Task, our model achieved a Pearson correlation of 0.7704 and ranked 10th, only 0.018 points behind the winner. An ablation study shows that all independent models contributed to the stacked model's performance, with the predictions from the BERT model fine-tuned in a multi-task fashion having the greatest impact in predicting lexical complexity. The code to reproduce our results is available in: https://github.com/kdrivas/lexical_complexity.

2 Background

The LCP Shared Task on SemEval-2021 asks participants to develop models that predict the com-

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 144–149
Bangkok, Thailand (online), August 5–6, 2021.

Sentence with Target	Complexity
His left hand is under my ***head****.*	0.125
Do therefore according to your wisdom, and don't let his gray ***head*** *go down to Sheol in peace.*	0.383

Table 1: Annotated sentences in the dataset of the LCP Shared Task. The target word is boldfaced.

plexity of a target word/MWE in a sentence in English (Shardlow et al., 2021). This Shared Task builds on previous editions that focused on Complex Word Identification (Paetzold and Specia, 2016; Štajner et al., 2018), with a key difference: complexity ratings are continuous scores instead of binary. Furthermore, the same target word/MWE can appear in more than one sentence but with different complexity scores. Table 1 presents an example from the data.

The data for the Shared Task is an extension of CompLex (Shardlow et al., 2020), a dataset with complexity ratings for target words/MWE in sentences in English in three domains: Bible, Europarl and Biomed. The dataset is split into two subs-tasks: LCP for single words and LCP for MWEs.

3 System Description

This section details our stacking approach to the LCP Shared Task Sub-task 1. An overview of our system can be seen in Figure 1.

3.1 Features

After joining all the data from both subtasks (single word and MWE), we extracted some features presented in (Yimam et al., 2018; Finnimore et al., 2019) and other custom ones, such as (1) the complexity of the target words in the lexicon proposed in (Maddela and Xu, 2018), (2) the predictions from four fine-tuned Transformer based models, and (3) the number of senses and dependencies of the target word/MWE.

3.1.1 Morphosyntactic and Lexical Features

First, we computed the number of characters and the number of words surrounding the target word/MWE. In addition, we obtained the part-of-speech of the first token and the syntactic dependencies of the whole target using the spaCy library.[1] We also counted the number of possible part-of-speech tags for the token using the Brown dictionary in NLTK.[2] Then, we counted the number of propositions, verbs, nouns, adverbs and got the ratio between the number of nouns and verbs using the whole sentence. Finally, we calculated the total number of syllables and morphemes.

3.1.2 N-gram Features

We formed n-grams considering one and two tokens surrounding the target word/MWE. Then, we computed their frequency in the Children's Book Test (Hill et al., 2015) and Simple Wikipedia (Kauchak, 2013). In addition, using the previous corpora, the Lang-8 corpus (Mizumoto et al., 2011) and the Tatoeba corpus,[3] we computed the frequency of the target tokens.

3.1.3 Word Complexity Lexicon

The lexicon created in (Maddela and Xu, 2018) contains complexity scores for more then 15,000 words. After lower-casing the words in the lexicon and the datasets from the Shared Task, we assigned the complexity from the lexicon to the words in the LCP data. If the word does not appear in the lexicon we assigned a null value.

3.1.4 Transformer-based Model Predictions

The last set of features is composed of the predictions of four pre-trained language models fine-tuned on the training data of both subtasks. The first three were a RoBERTa (Liu et al., 2019) and an XLNet (Yang et al., 2019) models that received as input the target word/MWE and a context window of 1, and a RoBERTa model with the target and a context window of 2. The last model was a BERT fine-tuned in a multi-task fashion with two tasks: LCP and Word Sense Disambiguation (WSD). For the former task, we only used the data generated with a window size of 1 and, for the latter, the Unified Evaluation Framework (Raganato et al., 2017).

Multi-Task Model. Given a sentence S of the dataset of the Shared Task and a complex word w in position a whose part of speech is p, we obtain a subsequence of size 1, $sub =< w_{a-1}, w_a, w_{a+1} >$; then:

$$CLS = BERT(sub) \quad (1)$$

where CLS is the CLS token of BERT, which

[1] https://spacy.io/

[2] https://www.nltk.org/

[3] Available in https://tatoeba.org/ under a CC-BY 2.0 FR licence.

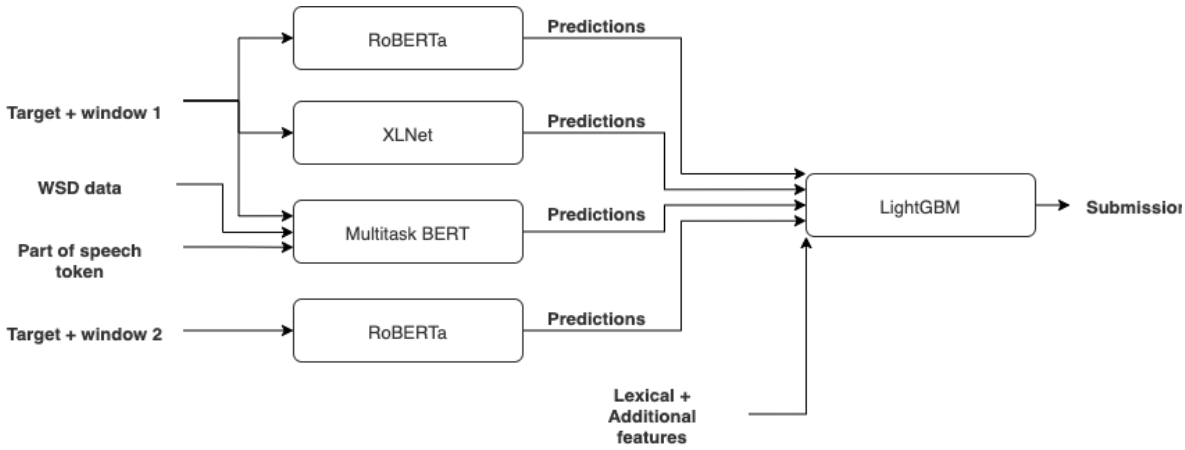

Figure 1: We used a LightGBM on the top of our architecture. It received the additional features and, the predictions from a XLNet and a BERT models using a window size of 1 and two RoBERTa models using a window size of 1 and 2.

represents the sentence. This representation is concatenated with the embedding token of p:

$$c = \text{concat}(CLS,\ \text{embed}\ (p)) \tag{2}$$

The concatenated vector is then used as input to a dropout layer and a linear layer:

$$out_1 = \ \text{Linear}\ (\ \text{Dropout}\ (c)) \tag{3}$$

Using out_1, we computed loss L_1 using mean squared error. After getting the first task loss, we computed the loss for the second one. Given an ambiguous sentence S and a sequence output of senses id A, we used the BertForTokenClassification implementation in HuggingFace[4] to obtain the output out_2, and then used cross entropy to compute loss L_2. Finally, we multiply a weight per each task loss to get the final overall loss:

$$L = W_1 * L_1 + W_2 * L_2 \tag{4}$$

Finally, we perform other experiments

3.2 Architecture

Our model architecture is shown in Figure 1. First, we got the predictions from the four language models. Then, we concatenated those predictions with the additional features, and stacked a LightGBM model that received them as input features.

4 Experimental Setup

As previously described, we used four different models: RoBERTa, XLNet, BERT and LightGBM. In addition, for training/fine-tuning each model we chose the Mean Absolute Error (MAE) as our validation metric.

4.1 RoBERTa and XLNet

We fine-tuned the models for 4 epochs with a batch size of 24. In addition, we used a learning rate of 2e-5 and Adam optimizer. We used the models for sequence classification provided by HuggingFace.[5]

4.2 Multitask BERT

We fine-tuned a BERT model using two tasks: LCP and WSD. We trained the WSD task using the Unified Evaluation Framework (Raganato et al., 2017), but filtered sentences with a size greater 22 tokens. For fine-tuning, we used a learning rate of 2e-5 and Adam optimizer. We fine-tuned the models for 5 epochs with a batch size of 32. We calculated the loss accumulating the gradients from both tasks. Also, we experimented with assigning different weights to each task, and found that the best configuration was 0.8 for LCP and 0.2 for WSD.

4.3 LightGBM

At the top of our architecture, we used a LightGBM model. Using Hyperopt, a bayesian optimization framework, we set up a max depth of 5, num-leaves

[4] https://huggingface.co/transformers/model_doc/bert.html#bertfortokenclassification

[5] https://huggingface.co/transformers/model_doc/roberta.html#robertaforsequenceclassification

of 8, min-sum-hessian-in-leaf of 0.9, a bagging-fraction of 0.9, a bagging-freq of 100, a learning-rate of 0.08, and a min-data-per-group of 100. We trained using 500 iterations with an early stopping of 90. Also, we declared the type of corpus and the part of speech as categorical features.

5 Results

The test set contains more than 1,000 sentences with 573 different target words. Table 2 shows the official evaluation metrics for each domain-corpus in the LCP dataset. Overall, we achieved a Pearson correlation of 0.7704, and finished in 10th place in the Shared Task Sub-task 1, only 0.018 points behind the winning submission.

Corpus	Pearson	Spearman	MAE	MSE
Bible	0.7536	0.7300	0.064	0.0074
Europarl	0.7492	0.7028	0.052	0.0045
Biomed	0.7898	0.7608	0.070	0.0083
Overall	0.7704	0.7361	0.618	0.0066

Table 2: Results in test set grouped by corpus domain.

The scores in the validation set (Table 3) follow a similar behaviour as those in the test set. For both, the corpus where our model achieves the best Pearson correlation is Biomed. However, looking at other metrics such as MAE, this corpus has the greatest error, with Europarl having the lowest. The differences may be because, even though the model may well capture the trend of the outputs, it could be more difficult to predict values in a corpus with higher variance of complexity scores, as is the case for Biomed (Figure 2).

Corpus	Pearson	Spearman	MAE	MSE
Bible	0.7353	0.6441	0.068	0.0072
Europarl	0.7946	0.7640	0.050	0.0039
Biomed	0.8571	0.8367	0.066	0.0075
Overall	0.8228	0.7643	0.062	0.0062

Table 3: Results in validation set grouped by corpus domain.

6 Ablation Study

Table 4 shows the contribution of each set of features (including predictions of fine-tuned models) to the final score. Although the predictions of

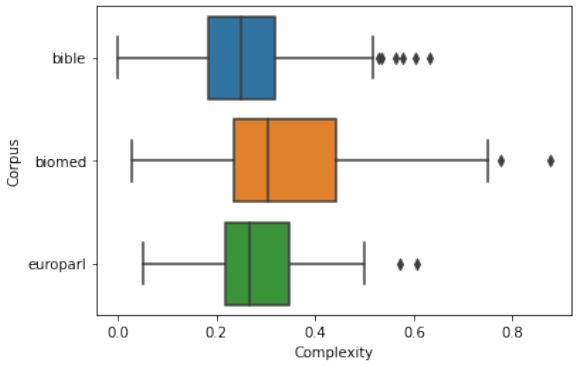

Figure 2: Distribution of the word complexity in validation set.

the fine-tuned Transformers-based models perform very well independently, the combination of all the predictions and the additional traditional features achieves the best performance in the validation set.

Another way of visualising the importance of each feature is using SHAP values (Lundberg and Lee, 2017). Figure 3 reports the 10 most important features for the LightGBM model, i.e. the impact of each feature in predicting the target complexity score. The X-axis shows the increase or decrease of target complexity, while the red and blue colours refer to the feature value's size. For example, in the case of feature `size_of_sentence`, if the number of characters is larger there will be a positive impact, i.e. the complexity will increase. On the other hand, if the sentence length is smaller, there will be a negative impact, i.e. the complexity will decrease. We can observe that the most important feature is the predictions given by the BERT Multitask model since they have the greatest impact. This signals that WSD data could benefit predicting lexical complexity. It is also noted that the predictions of the Transformers-based models are in the top 5 of importance. Other features, such as the size of the sentence or the number of word senses, also have good contributions to the impact.

7 Conclusion

In this paper, we presented our system for the single word complexity prediction sub-task in the LCP Shared Task. Our approach consisted of combining lexical features and predictions from fine-tuned pre-trained Transformer-based models. We found that each set of features achieved a good performance on their own, and that combining all of them achieved our best result. In particular, we found that fine-tuning a pre-trained Transformer-

Approach	Pearson	Spearman	MAE	MSE
(a) BERT multitask with a window of 1	0.7972	0.7457	0.0642	0.00691
(b) BERT with a window of 1	0.7936	0.7507	0.0650	0.00703
(c) RoBERTa with a window of 1	0.7760	0.6946	0.0691	0.00776
(d) RoBERTa with a window of 2	0.7902	0.7179	0.0659	0.00729
(e) XLNet with a window of 1	0.7761	0.7253	0.0704	0.00795
(f) LightGBM with additional features	0.7859	0.7326	0.0663	0.0073
(a), (c), (d), (e) and (f)	0.8228	0.7643	0.0616	0.00618

Table 4: Results of each approach on validation data

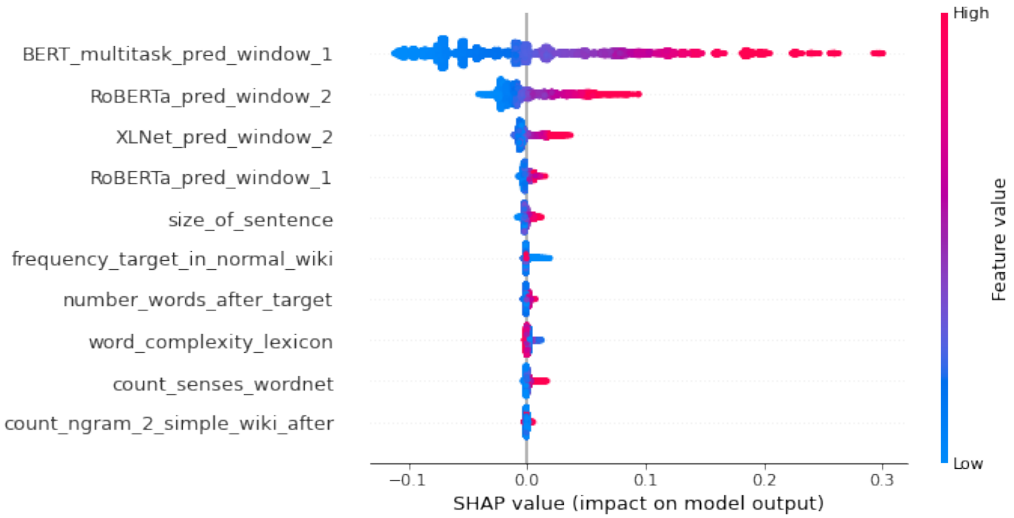

Figure 3: Shap analysis for the top 10 most important features

based model using multi-task learning with data from word sense disambiguation helped the most with learning to predict lexical complexity.

Considering that there were unseen tokens in validation and test sets, the task resembles a zero shot classification problem. Therefore, as future work, semi-supervised learning approaches or data augmentation algorithms could be explored, and training in a multitask fashion another transformer-based models like RoBERTa.

References

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.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple languages. 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 970–977, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3149–3157, Red Hook, NY, USA. Curran Associates Inc.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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)*.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

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 14th International Workshop on Semantic Evaluation (SemEval-2021)*.

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*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.