LangResearchLab_NC at SemEval-2021 Task 1: Linguistic Feature Based Modelling for Lexical Complexity

Raksha Agarwal, Niladri Chatterjee
Indian Institute of Technology Delhi
Hauz Khas, Delhi-110016, India
raksha.agarwal@maths.iitd.ac.in
niladri@maths.iitd.ac.in

Abstract

The present work aims at assigning a complexity score between 0 and 1 to a target word or phrase in a given sentence. For each Single Word Target, a Random Forest Regressor is trained on a feature set consisting of lexical, semantic, and syntactic information about the target. For each Multiword Target, a set of individual word features is taken along with single word complexities in the feature space. The system yielded the Pearson correlation of 0.7402 and 0.8244 on the test set for the Single and Multiword Targets, respectively.

1 Introduction

Presence of complex words can lead to poor comprehension of a text. Identification of such complex words in a given text is a core component in the task of Automatic Simplification and Evaluation (Shardlow, 2013). The Lexical Complexity Prediction Task of SemEval 2021 (Shardlow et al., 2021) aims at development of systems for prediction of complexity scores for a target word/phrase in a given sentence. In literature, binary classification of target words in a text into complex or non-complex is referred to as Complex Word Identification (CWI) (Paetzold and Specia, 2016; Zampieri et al., 2017; Gooding and Kochmar, 2018; AbuRa'ed and Saggion, 2018; Yimam et al., 2018). Unlike previous works, a continuous complexity score is assigned to the target word in the present task which is referred to as Lexical Complexity Prediction (LCP) (Shardlow et al., 2020). For the present work, regression is performed for LCP on a set of linguistic features covering semantic, syntactic and contextual aspects of the target word as described in Section 3. Additionally, various lexicon based features are used to indicate the rarity of target words. The system achieves 0.8194 Pearson correlation for Single Word Target and 0.7482 for Multiword Target on the trial set.

2 Task Setup

The task is divided into two subtasks, namely Single Word Target and Multiword Target based on the length of the target. The dataset and evaluation metrics are described below.

- **Dataset:** The dataset consists of an augmented version of CompLex (Shardlow et al., 2020). It comprises sentences from three corpora, viz. World English Bible Translation, English Portion of the European Parliament proceedings, and articles from CRAFT corpus belonging to biomedical domain. It is split into three subsets Train, Trial, and Test.
- **Evaluation Metrics:** The systems are evaluated using Pearson correlation coefficient (P), Spearman rank correlation coefficient (S), Mean absolute error (MAE) and Coefficient of Determination (R^2).

3 Features

In this section we present the details of the feature space used in the present work.

3.1 Corpus Features

A feature, named Corpus, is used to indicate to which of the 3 corpora the input sentence belongs.

3.2 Shallow Features

Word level shallow features used in the present work are number of letters (Nlet), syllables (Nsyl), vowels (Nvow), percentage of upper case alphabets (PerUp), simple universal part-of-speech tag (POS), and detailed Penn part-of-speech tag (Tag) of the target word extracted using SpaCy.

3.3 NLTK WordNet Features

Number of hypernyms (Nhyper) and number of morphemes (Nmorph) of the target word consider-

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 120–125
Bangkok, Thailand (online), August 5–6, 2021.

ing its POS tag in the given sentence are also used as features.

3.4 Exquisite Corpus (EC) Features

Exquisite Corpus[1] compiles texts from seven different domains namely Wikipedia, Subtitles, News, Books, Web, Twitter and Reddit. We have used the frequency (WordFreq) in EC and the Zipf frequency (ZipfFreq) of the target word as features (van Heuven et al., 2014).

3.5 SUBTLEX Features

Frequency (SubtFreq) of the target word extracted from SUBTLEXus[2] and its Contextual Diversity (ConDiversity) i.e. percent of the films in which the word appears are used as features.

3.6 Language Model (LM) Features

Given an input sentence $S = w_1 w_2 \dots w_N$ and a target word w_t where $t \in 1, 2, \dots N$, the following features are extracted from a trigram language model trained on the Gigaword corpus[3].

- Perplexity of the input sentence (Perplexity) computed as: $Perplexity(S) = \sqrt[N]{1/P(w_1 w_2 \dots w_N)}$
- The phrase score (PhrScore) of $w_j \dots w_t \dots w_k$ defined as $log_{10} P(w_j \dots w_t \dots w_k)$ where $j = max(1, t-2)$ and $k = min(N, t+2)$
- Average of conditional probabilities involving the target word (AvgCP)

$$Avg \begin{pmatrix} P(w_t \mid w_{t-1}, w_{t-2}), \\ P(w_{t+1} \mid w_t, w_{t-1}), \\ P(w_{t+2} \mid w_{t+1}, w_t) \end{pmatrix}$$

3.7 Character Language Model (CharLM) Feature

The probability of the target word (Prob3c) calculated using trigram character language model is considered as a feature. The trigram[4] probabilities are calculated using letter counts from Google Web Trillion Word Corpus. Suppose a word W consist of N letters, $W = w_1 \dots w_N$ then, the corresponding feature value will be computed as: $Prob3c(W) = \frac{1}{N-2} \sum_{i=1}^{N-2} log_{10} P(w_i w_{i+1} w_{i+2})$

[1] https://pypi.org/project/wordfreq/
[2] https://github.com/Wonderlic-AI/wonderlic_nlp
[3] lm_giga_64k_nvp_3gram.zip
[4] http://norvig.com/ngrams/count_3l.txt

3.8 Psycholinguistic Features

The following features are extracted using MRC psycholinguistic database (Wilson, 1988): Age of acquisition (AOA), Concreteness (CONC), Imageability (IMAG) and Meaningfulness ratings (MeanC, MeanP) of the target word .

3.9 Kucera and Francis (KF) Features

The features derived by Kučera and Francis (1967), namely target word's written frequency of occurrence (KFFreq) and the number of categories of text in which the target word was found (KFNcats) are used.

3.10 Ogden Feature

A binary feature is used to indicate presence of the target word in the list of 1000 words included in Ogden's Basic English[5] (IsOgden).

3.11 Inquirer Tag Features

The General Inquirer classifies about 7500 words using 182 General Inquirer categories developed for social science content analysis (Stone et al., 1966). A binary feature is created for each category to indicate its occurrence for the target word. The POS tag of the target is matched with the 'OthTags' category to filter out incompatible categories as given in Table 1

POS of the Target	Compatible OthTags
NOUN | PRON | PROPN	NOUN | PRON
VERB | AUX | ADV	VERB | SUPV

Table 1: Inquirer Tags Filtering

4 Single Word Target

In the Single Word Target task, complexity scores between 0 to 1 needs to be assigned for a target word of the input sentence. Various regression models are trained using the optimal set of features using scikit-learn[6]. The results are presented in Table 2. For both Decision Tree and Extra Tree Regressors the maximum depth (maxdepth) is tuned between 1 to 20, and the optimal maxdepth is found to be 6 and 8, respectively. Random Forest Regressors with the default setting produced the best results for the trial dataset. Using the above, our submission to the shared task achieved 0.7402 Pearson correlation on the test set.

[5] http://ogden.basic-english.org/
[6] https://scikit-learn.org/stable/

Regressor	P	S	MAE	R^2
Decision Tree	0.761	0.699	0.069	0.58
Extra Tree	0.757	0.650	0.071	0.57
Gradient Boosting	0.794	0.731	0.065	0.63
Random Forest	**0.819**	**0.748**	**0.062**	**0.69**
+Bagging	0.805	0.738	0.064	0.64
+Adaptive Boosting	0.798	0.734	0.065	0.63

Table 2: Results on the Trial Set

4.1 Feature Importance

The Gini importance of the top 5 features are reported in Table 3. Gini importance of a feature is computed as the (normalized) total reduction of the mean squared error brought by that feature. The importance of the features is also analyzed by removing a set of features at a time and training a Random Forest Regressor for the reduced feature space. Each of the features from the optimal feature space has a positive effect on the performance of the system as indicated in Table 4. The experiments indicate that exclusion of Exquisite Corpus features led to the maximum decline in the results. Hence, this may be considered as the most important feature subset.

Feature	Gini importance
ConDiversity	0.443
Prob3c	0.072
ZipfFreq	0.068
Perplexity	0.067
AvgCP	0.060

Table 3: Gini Importance of Features

4.1.1 Inquirer Tags Importance

The effect of inclusion of Inquirer Tags in the feature space has a positive effect however the magnitude is low. This may be due to the low coverage of these features as reported in Table 5. The coverage is defined as the percentage of target words having at least one Inquirer Tag.

4.2 Additional Features

The following set of features when included in the feature space led to a decrease in performance for the present task on the trial set.

- Etymological Feature: The ISO code of the target word's origin language

Features	P	S	MAE	R^2
All	**0.819**	**0.748**	**0.062**	**0.67**
w/o Ogden	0.816	0.744	0.063	0.66
w/o Inquirer	0.815	0.744	0.063	0.66
w/o KF	0.815	0.746	0.063	0.66
w/o WordNet	0.814	0.747	0.063	0.66
w/o Psych	0.813	0.740	0.063	0.66
w/o LM	0.810	0.744	0.063	0.65
w/o CharLM	0.806	0.747	0.064	0.65
w/o Corpus	0.798	0.740	0.065	0.63
w/o SUBTLEX	0.795	0.725	0.066	0.63
w/o Shallow	0.786	0.728	0.067	0.61
w/o EC	0.782	0.713	0.067	0.61

Table 4: Feature Set Elimination Results for the Trial Set

Data	All	Bible	Biomed	Europarl
Train	21.14	20.23	21.48	21.77
Trial	22.09	23.78	19.26	23.08
Test	23.77	19.79	27.34	24.06

Table 5: Inquirer Tags Coverage

- Named Entity Feature: The named entity tag[7] of the target word.

Post task evaluation on the test set indicates their inclusion improves the performance of the system. (See Table 6)

4.3 BERT Features

BERT was introduced in (Devlin et al., 2019), and its usage has resulted in state-of-the-art performance for various downstream NLP tasks, such as Question Answering, Textual Entailment and Paraphrase detection. In the present work, BERT embedding for the target word is extracted from the pre-trained BERT-base-uncased model[8]. Additionally, in an effort to enhance the contextualized BERT embeddings (Agarwal et al., 2020), the embedding vector is supplemented with the feature vector corresponding to linguistic features described in Section 3. Finally, a Neural Network is trained to minimize the Mean Absolute Error using Adam optimizer (Kingma and Ba, 2015). Hyper parameter tuning is performed using hyperas[9] and TPE algorithm. The number of intermediate dense layers are tuned between {2, 3}. The encoding dimensions are tuned between {50, 100,

[7] extracted using https://spacy.io/api/entityrecognizer
[8] uncased_L-12_H-768_A-12.zip
[9] https://pypi.org/project/hyperas/

Included in Feature Space		Trial				Test			
Etymology	**NamedEntity**	**P**	**S**	**MAE**	**R^2**	**P**	**S**	**MAE**	**R^2**
No	No	**0.8194**	**0.7478**	**0.0624**	**0.6681**	0.7402	0.7005	0.0661	0.5440
Yes	No	0.8115	0.7451	0.0633	0.6565	**0.7421**	**0.7013**	**0.0660**	**0.5486**
No	Yes	0.8113	0.7440	0.0631	0.6561	0.7404	0.6966	0.0661	0.5464
Yes	Yes	0.8175	0.7466	0.0627	0.6654	0.7418	0.6974	0.0659	0.5475

Table 6: Results for Additional Features

200, 300, 500, 700, 1000} and dropouts between {0.1, ..., 0.9}. Batch size is set to 16. The results are presented in Table 7. It can be observed that BERT embeddings do not improve the performance. Moreover, Neural Networks when applied on just linguistic features have a lower performance than Random Forest Regressors.

4.4 Error Analysis

Error analysis indicates that absolute error for 87% test samples were less than 0.10. Samples belonging to Biomedical corpus had highest errors. Some predictions of the proposed model are presented in Table 8. The correlation between the actual and predicted complexity for similar targets in dissimilar contexts is high. However, it is revealed that difference in complexity of proper noun targets in distinct contexts could not be captured effectively through the present set of linguistic features.

5 Multiword Target

In the present task the Multiword Targets are pairs of two adjacent words. We have experimented with two approaches for predicting complexity scores for Multiword Targets, as described in Section 5.1 and Section 5.2

5.1 Single Word Combination

In this approach, each word of a Multiword target is considered as individual single word targets, and the complexity scores are predicted using the Single Word Target[10] model. The individual word scores are combined using Average, Maximum, and Minimum. Additionally, Algebraic Sum $(a + b - ab)$ and Product (ab) of the individual scores are also considered. These are taken from Fuzzy s-norm and t-norm (Klir and Yuan, 1995). The results are indicated in Table 9. For both trial and test set, maximum of the complexity score of each word of the multiword target gives the least MAE and the highest R^2 value. But, the highest P for trial set is obtained when algebraic sum of the individual complexity scores are taken and highest S is obtained when product of the individual complexity scores are taken. For the test set, algebraic sum gives highest P and S.

[10]Random Forest Regressor w/o additional features

5.2 Feature Combination

In this approach features corresponding to the individual words are concatenated, and then a regression model is trained with the increased feature space for complexity prediction. The individual target word complexity value predicted by the Single Word Target model is also considered as a feature. The results are presented in Table 10. Bagging and Adaptive Boosting are applied on Random Forest. The results indicate that inclusion of individual complexity scores enhances the performance of the system, and the best results are obtained for Bagging ensemble. Our submission to the shared task was derived using Bagging on the Random Forest Regressor. The feature set contains individual word features along with complexity scores. It achieved Pearson correlation of 0.8244 on the test set.

6 Conclusion

Identification of difficult words is an important task for Automatic Text Simplification. LCP aims at assigning scores to words of a given sentence to indicate its complexity. In this work we utilize word level features to capture its lexical, semantic and syntactic information. LM based features are used for indicating the semantics of the target word in a given context. Frequency and occurrence based features are used to indicate the overall rarity of the target words. For Single Word Target, Random Forest Regressors trained on the linguistic feature set achieved the highest results. Error analysis revealed that the model can be further improved to capture the context of the target word.

For Multiword Target, two approaches were explored. In the first approach complexity scores of individual target words predicted by the Sin-

Feature Set	# Dense Layers	Dimension	Dropouts	P	S	MAE	R^2
Linguistic	2	50,200	0.1, 0.3	0.752	0.698	0.070	0.563
BERT	3	300,300,1000	0.3,0.1,0.1	0.732	0.678	0.071	0.532
BERT + Linguistic	2	300,200	0.1,0.1	0.714	0.660	0.072	0.502

Table 7: Results on Trial Set for Neural Network

Input Sentence	Target Word	Actual Complexity	Predicted Complexity
Saul arose, and they went out both of them, he and Samuel, abroad.	Saul	0.3676	0.3398
Saul said to his servants, "Provide me now a man who can play well, and bring him to me."	Saul	0.3529	0.3383
Samuel said to Saul, "Why have you disturbed me, to bring me up?"	Saul	0.2778	0.3303
These results, as well as this study, suggest that a considerable amount of maternal cholesterol can be transferred to the murine fetus.	amount	0.2031	0.2048
This wild-type staining pattern may simply reflect the fact that decreasing the amount of mutant protein by half makes it undetectable by immunocytochemistry.	amount	0.2375	0.2207

Table 8: System Predictions

Combination Strategy	Trial				Test			
	P	S	MAE	R^2	P	S	MAE	R^2
Average	0.7329	0.7239	0.1220	0.0437	0.8098	0.8101	0.1314	0.0110
Maximum	0.6872	0.6733	**0.1021**	**0.2861**	0.7907	0.7916	**0.1041**	**0.3433**
Minimum	0.6964	0.6970	0.1534	-0.4056	0.7036	0.7064	0.1648	-0.5466
AlgebraicSum	**0.7391**	0.7217	0.1270	0.0598	**0.8193**	**0.8104**	0.1049	0.3349
Product	0.7047	**0.7298**	0.3153	-3.8253	0.7704	0.8063	0.3257	-3.9447

Table 9: Results for Multiword Target for Single Word Combination

Individual Complexity Predictions as a feature	Regressor	P	S	MAE	R^2
No	Random Forest	0.7327	0.7253	0.0885	0.5110
	+Bagging	0.7299	0.7294	0.0877	0.5118
	+Adaptive Boosting	0.7386	0.7369	0.0880	0.5167
Yes	Random Forest	0.7234	0.7256	0.0872	0.5134
	+Bagging	**0.7482**	**0.7510**	**0.0830**	**0.5517**
	+Adaptive Boosting	0.7455	0.7427	0.0853	0.5408

Table 10: : Results for Multiword Target on the Trial Set using Feature Combination

gle Word model were combined using different strategies, while in the second, the feature space was expanded to accommodate features and complexity scores corresponding to individual target words. The latter yielded the best results. Our system achieved 36^{th} and 17^{th} rank with respect to the two subtasks. The difference in the correlation value between the top performer is less than 0.05 for Single Word Target and 0.04 for Multiword Target.

Acknowledgments

Raksha Agarwal acknowledges Council of Scientific and Industrial Research (CSIR), India for supporting the research under Grant no: SPM-06/086(0267)/2018-EMR-I

References

Ahmed AbuRa'ed and Horacio Saggion. 2018. LaSTUS/TALN at complex word identification (CWI) 2018 shared task. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–165, New Orleans, Louisiana. Association for Computational Linguistics.

Raksha Agarwal, Ishaan Verma, and Niladri Chatterjee. 2020. LangResearchLab_NC at FinCausal 2020, task 1: A knowledge induced neural net for causality detection. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 33–39, Barcelona, Spain (Online). COLING.

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.

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@NAACL-HLT 2018, New Orleans, LA, USA, June 5, 2018*, pages 184–194. Association for Computational Linguistics.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190. PMID: 24417251.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

George Klir and Bo Yuan. 1995. *Fuzzy sets and fuzzy logic*, volume 4. Prentice hall New Jersey.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. University Press of New England.

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)*.

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, Sofia, Bulgaria. Association for Computational Linguistics.

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)*.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. *The general inquirer: A computer approach to content analysis.* MIT press.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

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.

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.