LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction

Abhinandan Desai, Kai North, Marcos Zampieri, Christopher Homan
Rochester Institute of Technology
Rochester, NY, USA
ad2724@rit.edu, kn1473@rit.edu, mazgla@rit.edu, cmh@cs.rit.edu

Abstract

This paper describes team LCP-RIT's submission to the SemEval-2021 Task 1: Lexical Complexity Prediction (LCP). The task organizers provided participants with an augmented version of CompLex (Shardlow et al., 2020), an English multi-domain dataset in which words in context were annotated with respect to their complexity using a five point Likert scale. Our system uses logistic regression and a wide range of linguistic features (e.g. psycholinguistic features, n-grams, word frequency, POS tags) to predict the complexity of single words in this dataset. We analyze the impact of different linguistic features on the classification performance and we evaluate the results in terms of mean absolute error, mean squared error, Pearson correlation, and Spearman correlation.

1 Introduction

Lexical complexity prediction (LCP) is the task of predicting the complexity value of a target word within a given text (Shardlow et al., 2020). Complexity within LCP is used as a "synonym for difficulty" (Malmasi and Zampieri, 2016)[1]. A complex word is therefore a word that a target population may find difficult to understand. Various LCP systems have been designed to identify words that may be found to be complex for children (Kajiwara et al., 2013), language learners (Malmasi et al., 2016), or people suffering from a reading disability, such as dyslexia (Rello et al., 2013). These systems have been utilized within assistive language technologies, lexical simplification systems, and in a variety of other applications.

LCP is related to complex word identification (CWI) (Paetzold and Specia, 2016). CWI is modeled as a binary classification task by assigning each target word with a complex or non-complex label. The shortcomings of modeling lexical complexity using binary labels have been discussed in previous work (Zampieri et al., 2017; Maddela and Xu, 2018), motivating the organization of SemEval-2021 Task 1: Lexical Complexity Prediction.[2] LCP models complexity in a continuum and the goal is to predict a target word's degree of complexity by assigning it a value between 0 and 1. This value may then correspond to one of the following labels: very easy (0), easy (0-0.25), neutral (0.25-0.5), difficult (0.5-0.75), or very difficult (0.75-1) (Shardlow et al., 2020).

In this paper, we describe (in detail in Section 4) the LCP-RIT entry to SemEval-2021 Task 1. We approached LCP from a feature engineering perspective with a particular focus on the adoption of psycholinguistic features, such as average age-of-acquisition (AoA), familiarity, prevalence, concreteness, and arousal, alongside the use of prior complexity labels. Our submitted system utilized a combination of these linguistic features, which we compared to a baseline model that only used statistical features: word length, word frequency and syllable count (Quijada and Medero, 2016; Mukherjee et al., 2016). On our training dataset, our submitted system achieved a mean absolute error (MAE) of 0.067, mean squared error (MSE) of 0.007, Person Correlation (R) score of 0.779, and a Spearman Correlation (ρ) score of 0.724. This surpassed our baseline model's performance by a MAE of 0.008, MSE of 0.003, as well as R and ρ scores of 0.075 and 0.062 respectively.

2 Related Work

Before SemEval-2021 Task 1: LCP, two CWI shared tasks were organized at one SemEval-2016 and the other at BEA-2018 (Paetzold and Specia, 2016; Yimam et al., 2018). While the first CWI provided participants with an English dataset, the

---

[1] The term "complex" within LCP is not necessarily related to the terms *simplex* and *complex* used in morphology.

second provided a multilingual dataset. The systems submitted to the English track of the second shared task (Yimam et al., 2018) performed better overall than the previous task (Paetzold and Specia, 2016), probably due to the properties of the two datasets (Zampieri et al., 2017). State-of-the-art neural net models and word embedding models performed worse than conventional models such as decision trees (DTs) and random forests (RFs) (Yimam et al., 2018). Among the conventional models, the use of statistical, character n-gram, and psycholinguistic features was found to be highly effective in improving CWI performance (Malmasi et al., 2016; Zampieri et al., 2016; Paetzold and Specia, 2016; Yimam et al., 2018).

Among the best performing systems in CWI 2018, Gooding and Kochmar (2018) used an ensemble of classifiers. They found that during their system's development, the boosting classifier AdaBoost, a random forest classifier, or a combination of both classifiers achieved the highest performance. These systems used multiple features such as the word's grammatical category, Google character n-gram frequency as well as a range of psycholinguistic features (Gooding and Kochmar, 2018).

Of the remaining systems, Aroyehun et al. (2018) and Hartmann and Borges dos Santos (2018) utilized statistical features, such as word length and number of syllables, psycholinguistic features such as familiarity, age of acquisition, concreteness, and imagery scores, and word n-grams. Hartmann and Borges dos Santos (2018) compared the performance of tree ensembles to a convolutional neural network (CNN). They found that their tree ensembles performed better than their CNN, especially when the target expression contained more than three words (Aroyehun et al., 2018).

3 Task and Dataset

The LCP shared task organizers provided participants with the CompLex corpus, an English multi-domain dataset with sentences from the Bible, the European Parliament proceedings, and a collection of biomedical texts. A pool of annotators, using a five point Likert scale, labeled the complexity of single words and multi-word expressions in CompLex (Shardlow et al., 2020).

Taking advantage of the annotation of single words and multi-word expressions, the LCP shared task was divided into two sub-tasks as follows:

- **Sub-task 1:** predicting the complexity score for single words;

- **Sub-task 2:** predicting the complexity score for multi-word expressions.

We chose to participate in sub-task 1. Sub-Task 1's training dataset contained 7,662 instances with its test dataset having 917 instances. 20% of the training dataset was used to test our system's performance during development. Sub-Task 1 received 54 system submissions.

4 System Overview

4.1 Model

Taking inspiration from the CWI systems discussed in Section 2, we adopted a random forest regressor (RFR) to predict the complexity values of each word within the test dataset. To achieve this, we tested the impact a variety of linguistic features have on LCP performance during our system's development. The RFR was taken from scikit-learn's ensemble module (Pedregosa et al., 2011). The RFR used a maximum of 120 trees and 750 features.

4.2 Features

We constructed a baseline RFR using the following statistical features and character trigrams. We then used psycholinguistic and additional features to see whether its baseline performance could be improved.

Statistical Features include word length, word frequency and syllable count. Zipf's Law implies that words that appear less frequently within a text are likely to be longer and therefore may be considered more complex than words that are more frequent and shorter (Quijada and Medero, 2016). In addition, words with a high number of syllables are difficult to pronounce and are subsequently hard to read (Mukherjee et al., 2016). As such, word length, word frequency and syllable count were considered to be good baseline statistical indicators of a word's complexity value.

Character N-grams include the use of character bigram and trigram frequencies. These frequencies were calculated by counting each bigram's and trigram's presence in the target words provided in Sub-Task 1's training dataset. Experimentation with bigrams and trigrams, along with a combination of both, found that the use of trigrams on their own was superior. This together with their use

within prior CWI systems justified their inclusion within our baseline model (Yimam et al., 2018).

Psycholinguistic Features include average age of acquisition (AoA), concreteness, familiarity, prevalence and arousal. AoA is the age at which a word's meaning is first learned. Concreteness refers to "the degree to which the concept denoted by a word refers to a perceptible entity" (Brysbaert et al., 2013). Familiarity and prevalence are somewhat similar. Familiarity is how well known the word is to an individual and was obtained through self-report (Gilhooly and Logie, 1980). Prevalence was calculated in accordance to the percentage of people who knew the word (Brysbaert et al., 2019). Lastly, arousal is a measure of how active or passive a word's meaning is interpreted as being[3]. For instance, the word "*nervous*" indicates more arousal than "*lazy*" (Mohammad, 2018). As such, grammatical categories such as adjectives, verbs, and adverbs may incite higher levels of arousal than nouns.

Average AoA was calculated by averaging the AoAs provided in the Living Word Vocabulary Dataset (Dale and O'Rourke, 1981) with an updated version of this dataset (Brysbaert and Biemiller, 2017). Both datasets consisted of AoA values for 44,000 English word meanings. Concreteness, familiarity and arousal values were taken from the MRC Psycholinguistic Database (Wilson, 1988) as well as three newer datasets each containing 37,058, 61,858 and 20,000 English words (Brysbaert et al., 2013, 2019; Mohammad, 2018).

Additional Features include part-of-speech (POS) tags as well as prior complexity labels. POS tags were generated by using the Python Natural Language Toolkit (Bird et al., 2009). Prior complexity labels were taken from the previous CWI shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018) and the Word Complexity Lexicon (Maddela and Xu, 2018). A combined dataset was then created that contained a total of 26,088 English words each with a binary complexity value.

5 Evaluation

5.1 Features

To determine the effect each feature had on our baseline model's performance, we used the fol-

lowing scores: mean absolute error (MAE), mean squared error (MSE), Pearson Correlation (R) and Spearman Correlation (ρ). Table 1 depicts each feature's performance on the training dataset. These criteria have also been used in the SemEval LCP test set evaluation.

Average AoA decreased the baseline model's MAE and MSE by 0.004 and 0.001 respectively. It likewise increased its R and ρ scores by 0.039. This generated the second highest R and ρ scores of 0.743 and 0.701 respectively. Average AoA is therefore a useful feature for LCP.

Brysbaert et al.'s prevalence and concreteness (Brysbaert et al., 2013, 2019) were also seen to improve the baseline model's performance with prevalence being the most notable. Prevalence decreased baseline MAE and MSE scores by 0.005 and 0.002 respectively. It also surpassed baseline R by 0.054 and ρ by 0.047, yielding the highest increases among all features. Concreteness (Brysbaert et al., 2013) also caused a slight increase in the baseline model's scores, being greater than that caused by MRC concreteness. Concreteness values (Brysbaert et al., 2013) increased the baseline model's R and ρ scores by 0.032 and 0.024 respectively, whereas the MRC concreteness values resulted in a slightly less impressive increase of 0.019 in both its R and ρ scores. However, there was little-to-no difference in MAE and MSE produced by either set of concreteness values.

	Performance			
Features	R	ρ	MAE	MSE
Baseline Features	0.704	0.662	0.075	0.010
Average AoAs	0.743	0.701	0.071	0.009
Prevalence	**0.758**	**0.709**	**0.070**	**0.008**
MRC Familiarity	0.727	0.687	0.073	0.009
Concreteness	0.736	0.686	0.072	0.009
MRC Concreteness	0.723	0.681	0.073	0.009
Arousal	0.722	0.676	0.074	0.009
POS Tags	0.701	0.663	0.075	0.010
Complexity Labels	0.727	0.686	0.072	0.009

Table 1: Feature performance on training dataset. The baseline model uses the statistical features and character trigrams. Best results in bold.

Two possible conclusions can be drawn: 1). The difference in the calculation of prevalence versus that of familiarity likely causes prevalence to be a greater indicator of word complexity[4], and 2). The

---

[3]The terms "active" and "passive" do not refer to the use of *active* or *passive voice* but rather the emotional or physical intensity associated with a word's meaning (Mohammad, 2018).

[4]Prevalence being the percentage of people who. know the word (Brysbaert et al., 2019). Familiarity being a self-reported

superior coverage of Brysbeart et al.'s prevalence (2018) and concreteness (2013) datasets (Brysbaert et al., 2019, 2013): being 52.62% and 57.51% respectively, compared to that of the MRC Psycholinguistic Database (Wilson, 1988): being 23.44%, suggests that there now exists larger and more up-to-date psycholinguistic datasets that are more useful for LCP feature engineering.

Arousal has never before been used for LCP or CWI. Due to its ability to differentiate grammatical categories, such as nouns and verbs, along with its ability to signify a word's intensity, we had speculated that arousal would be able to help predict a word's complexity. Arousal was found to have no significant effect on the baseline model's performance. Nevertheless, once added to our submitted system, it slightly decreased its MSE by 0.001 and increased its R and ρ scores by 0.002 and 0.001 respectively.

POS tags was the worst performing feature as POS tags had little affect on improving our model's performance. It achieved the same MAE and MSE values as our baseline model: 0.075 and 0.01 respectively. Regarding R score, only a slight increase of 0.001 was observed. POS tags was the only feature that saw a decrease in our model's ρ score, worsening its performance by 0.003. This suggests that a word's grammatical category may not impact its degree of complexity. This is also supported by Arousal's lack of improved performance.

Given that prior complexity labels are directly related to complexity prediction, it was believed that they would be the most influential in improving overall performance. Instead, AoA, prevalence and concreteness were all found to be more beneficial with higher or identical MAE, MSE, ρ, and R scores. Complexity labels only saw a slight decrease in MAE and MSE by 0.003 and 0.001 respectively and a slight increase in ρ and R scores by 0.023 and 0.024 respectively. The binary nature of prior CWI datasets is likely responsible for this phenomenon, as binary 0 or 1 complexity values are not well suited for a regression-based task, such as LCP. This would have resulted in the same problem faced by previous CWI systems: the misclassification of words on the decision boundary.

measure of an individual's awareness of the word (Gilhooly and Logie, 1980).

5.2 Models

The results for the three models on the training dataset are presented in Table 2. This is then followed by a short description of each model as well as our performance on the test dataset.

	Performance			
Model	R	ρ	MAE	MSE
Model 1	0.772	0.717	0.068	0.008
Model 2	0.777	0.724	0.067	0.008
LCP-RIT	**0.779**	**0.724**	**0.067**	**0.007**

Table 2: Model performance on training dataset.

Model 1 - Top 3 Features: Adding the top 3 features of average AoA, prevalence and concreteness to our baseline model reduced its MAE and MSE by 0.007 and 0.002 respectively and increased its R score by 0.068 and its ρ score by 0.055. It attained a new MAE of 0.068 which was noticeably better than our baseline model's previous MAE of 0.075. This goes to the show that inclusion of psycholinguistic features has a positive impact on the performance of an LCP system.

Model 2 - Top 5 Features: A small improvement was seen after adding the fourth and fifth best performing features to Model 1, namely, MRC familiarity and prior complexity labels. Model 2 increased Model 1's R and ρ scores by 0.005 and 0.007. However, it failed to improve Model 1's MAE or MSE. This small increase in performance was due to the prior top 3 features of average AoA, prevalence and concreteness already having captured those instances caught by MRC familiarity and prior complexity labels. This further proves the redundancy of the MRC Psycholinguistic Database (Wilson, 1988) as well as binary complexity labels for LCP feature engineering.

LCP-RIT: Our final model submitted to the official evaluation used the psycholinguistic features of average AoA, prevalence, concreteness and arousal together with our baseline model's features of word length, syllable count, word frequency and character trigrams to predict the lexical complexity of single words. On the training dataset of SemEval-2021 Task 1: LCP, we achieved a MAE of 0.067, MSE of 0.007, R score of 0.779, and ρ score 0.724. We performed less well on the single word test dataset with an MAE and MSE of 0.072 and 0.009 respectively and ρ and R scores of 0.709 and 0.653 respectively. This reduced performance may be

indicative of our submitted system being overfit on our training dataset.

6 Conclusion

We carried out multiple experiments evaluating the impact of linguistics features in LCP using the CompLex dataset for English. We have shown that several psycholinguistic features help with LCP. Average AoA, prevalence and concreteness were all found to be beneficial, whereas MRC familiarly, MRC concreteness and prior complexity labels were proven to be redundant. We would like to explore other features described in Shardlow et al. (2021). In terms of performance, we believe that the multiple features we tested allowed us to get close to the maximum performance for this dataset using regression. A possible alternative for better performance is to test state-of-the-art transformer models. Furthermore, we are interested in looking at the performance of these features for LCP in languages other than English and for multilingual datasets.

Acknowledgments

We would like to thank the LCP shared task organizers for proposing this interesting shared task and for making the data available.

References

Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, and Alexander Gelbukh. 2018. Complex Word Identification: Convolutional Neural Network vs. Feature Engineering . In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Marc Brysbaert and Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand english word meanings. *Behavioural Research*, 49:1520–1523.

Marc Brysbaert, Pawel Mandera, Samantha McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior Research Methods*, 51:467–479.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Edgar Dale and Joseph O'Rourke. 1981. *The living word vocabulary, the words we know: A national vocabulary inventory*. World Book.

Ken Gilhooly and Robert Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.

Nathan Hartmann and Leandro Borges dos Santos. 2018. NILC at CWI 2018: Exploring Feature Engineering and Feature Learning. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73.

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.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2016. MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifier. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia. Association for Computational Linguistics.

Niloy Mukherjee, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. JU_NLP at SemEval-2016 Task 11: Identifying Complex Words

in a Sentence. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California. Association for Computational Linguistics.

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)*, San Diego, California. Association for Computational Linguistics.

Febian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maury Quijada and Julie Medero. 2016. HMC at SemEval-2016 Task 11: Identifying Complex Words Using Depth-limited Decision Trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California. Association for Computational Linguistics.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proceedings of the INTERACT 2013: 14th IFIP TC13 Conference on Human-Computer Interaction. Cape Town, South Africa, 2013*. INTERACT.

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)*, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2021. Predicting lexical complexity in english texts. *arXiv preprint arXiv:2102.08773*.

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, Luci Specia, Sanja Štajner, Anaïs Tack, 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*, New Orleans, United States. 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)*, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for ComplexWord Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California. Association for Computational Linguistics.