CLULEX at SemEval-2021 Task 1: A Simple System Goes a Long Way

Greta Smolenska, Peter Kolb, Sinan Tang,
Mironas Bitinis, Héctor Hernández, Elin Asklöv

Babbel|Lesson Nine GmbH
{gsmolenska, pkolb, stang, mbitinis, hhernandez, caskloev}@babbel.com

Abstract

This paper presents the system we submitted to the first Lexical Complexity Prediction (LCP) Shared Task 2021. The Shared Task provides participants with a new English dataset that includes context of the target word. We participate in the single-word complexity prediction sub-task and focus on feature engineering. Our best system is trained on linguistic features and word embeddings (Pearson's score of 0.7942). We demonstrate, however, that a simpler feature set achieves comparable results and submit a model trained on 36 linguistic features (Pearson's score of 0.7925).

1 Introduction

Lexical complexity relates to complexity of words. Its assessment can be beneficial in a number of fields, ranging from education to communication. For instance, lexical complexity studies can assist in providing language learners with learning materials suitable for their proficiency level or aid in text simplification (Siddharthan, 2014). These studies are also a central part of reading comprehension, as lexical complexity can predict which words might be difficult to understand and could hinder the readability of the text. Lexical complexity studies typically make use of Natural Language Processing and Machine Learning methods (Paetzold and Specia, 2016).

Previous similar studies focus on Complex Word Identification (CWI), which is a process of identifying complex words in a text (Shardlow, 2013). In this case, lexical complexity is assumed to be binary - words are either complex or not. LCP Shared Task 2021 addresses this limitation by introducing a new dataset designed for *continuous* rather than *binary* complexity prediction (Shardlow et al., 2021).

In this paper, we describe a *single-word* lexical complexity prediction system. Our goal is to demonstrate that a simple system can achieve results comparable to more complex ones. Therefore, we focus on feature engineering rather than model tuning.

2 Related Work

2.1 Lexical Complexity

Over the years, studies on lexical complexity have ranged from research on the overall *readability* enhancement and *text simplification* to studies focusing specifically on lexical complexity.

Some of the earlier work on lexical complexity targeted communication enhancement of medical documents by assessing the familiarity of medical terminology (Zeng et al., 2005). Paetzold and Specia (2013) showed that the absence of lexical simplification in Automatic Text Simplification (ATS) systems yielded texts that readers might still find too complex to understand.

CWI has then gained more interest, and two Shared Tasks have been organised with the goal of establishing state-of-the-art performance in the field. SemEval-2016 Task 11 approached CWI as a *binary* classification task and collected a dataset for English which was annotated by non-native speakers (Paetzold and Specia, 2016). Zampieri et al. (2017) showed that such data annotation approach was not optimal. The second Shared Task addressed the limitations by introducing a multilingual dataset for Spanish, German, English and French and approaching the problem as both, a *binary* and a *probabilistic* complexity prediction task (Štajner et al., 2018).

2.2 Feature and Model Selection

In lexical complexity prediction tasks, linguistic features and word frequency measures have been proven to be among the most effective features. The winning systems developed for the CWI 2018

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 632–639
Bangkok, Thailand (online), August 5–6, 2021.

Shared Task (Yimam et al., 2018) use various lexical features, such as word N-gram, POS tags, and syntactic dependency parse relations. Moreover, they also include different variants of word frequency features, CEFR levels, and a few more.

As for the choice of algorithms, Gooding and Kochmar (2018) has achieved the best performing systems in English monolingual tasks using classifiers with ensemble techniques, such as `AdaBoost` with 5000 estimators and the aggregation classifier of `Random Forest`. The winning systems for multilingual tracks (Kajiwara and Komachi, 2018) also employ random forest models.

3 LCP Shared Task 2021 Setup

The LCP Shared Task 2021 aims to predict the complexity value of words in their context. It is divided into two sub-tasks: predicting the complexity score of 1) *single words* and 2) *multi-word expressions*. In this paper, we present a system for the first sub-task.

The Shared Task uses the CompLex corpus (Shardlow et al., 2020). In addition to the target word, it includes contextual information which is represented by a sentence where the word appears and its source or domain: *Bible* (Christodouloupoulos and Steedman, 2015), *Europarl* (Koehn, 2005) or *biomedical texts* (Bada et al., 2012). Each word in the dataset is evaluated by around 7 annotators from English speaking countries. The complexity labels are based on a 5-point Likert scale scheme (*very easy* to *very difficult*). The final dataset consists of 7,662 training and 917 testing instances.

The Shared Task baseline system uses a `linear regression` model. It is trained on log relative frequency and word length features, resulting in a Mean Absolute Error (MAE) of **0.0867**.

4 Methodology

In this section, we describe the methodology that we follow in the design of our system, including the used data, feature engineering and the training steps. The study relies on an in-depth experimentation with features. We aim to find out which linguistic information is the best predictor of lexical complexity.

4.1 Data Collection

For the computation of some features, we use additional data sources. We extract word frequencies from nine corpora that cover different domains and complexity levels: *BNC corpus*[1], *Simple Wikipedia and English Wikipedia*[2], *SubIMDB*[3] and English monolingual corpora from the *OPUS* project[4]: *bible-uedin*[5], *EMEA*[6], *Europarl*[7], *News-Commentary*[8] and *OpenSubtitles* 2018[9]. We additionally use two word lists with annotated CEFR levels (Common European Framework of Reference for Languages, which organises language proficiency in six levels, A1 to C2)[10] and the *Age of Acquisition* dataset[11].

4.2 Features

We consider a) word and sentence-level features (or *linguistic* features), b) *frequency* features and c) *word embeddings*.

On a word level, we compute the linguistic information, i.e. *character*, *syllable* and *phoneme counts*, *universal part-of-speech* tag and *named entity* tag (extracted with Stanza NLP toolkit) (Qi et al., 2020). We also compute scores that pertain to language learning such as *age of acquisition*, *percentage of population that knows the word* and *word prevalence* (Kuperman et al., 2012). Finally, we use two CEFR word lists and split them into five subsets each (one per CEFR level). Each word is assigned a *boolean value* depending whether it appears in one of the subsets.

On a sentence level, lexical complexity is represented by *lexical diversity rate* (*unique* words divided by *all* words). Syntactic complexity and readability are represented by the *average sentence length* and the *Linsear Write* score, which is a readability measure used to assess the difficulty of U.S. military manuals (Klare, 1974). We also make special use of the *OpenSubtitles* frequencies: *vocabulary percentage per CEFR level* is computed by splitting the corpus into five subsets and represents the distribution of words among the five frequency ranges; *difficult word percentage* relates to words containing two and more syllables that do not appear in top 200 most common words in the corpus; *unknown word percentage* represents

[1] BNC
[2] Wikipedia Monolingual Corpora
[3] SubIMDB
[4] OPUS resources
[5] bible-uedin
[6] EMEA
[7] Europarl
[8] News-Commentary
[9] OpenSubtitles 2018
[10] The Oxford 5000 and Kelly list for English
[11] AoA

the percentage of words that do not appear in the corpus at all. The final *text complexity score* is a normalised sum of all sentence-level scores.

Additionally, we calculate different types of *frequencies*, i.e. log relative, absolute (raw), frequency rank (word rank in a frequency list) and ZIPF frequency (Zipf, 1949), from the nine corpora.

Finally, we experiment with pre-trained *word embeddings*, including fastText for English and BERT's embeddings (Mikolov et al., 2018; Devlin et al., 2018). However, we ablate fastText word embeddings from the final feature set as they slightly degrade the overall performance.

4.3 Training, Tuning & Testing

The focus of our study is to achieve the best results through feature engineering rather than model hyperparameter tuning. During all experiments, we utilise the open source Machine Learning software WEKA (Frank et al., 2016) with the default algorithm hyperparameter settings and apply 10-fold cross-validation.

4.3.1 Models

First, we select several Machine Learning algorithms for further experiments with the features. During this step, we use *word* and *sentence-level* features with a subset of *frequency* features.

Due to the nature of the dataset target values, we employ classifiers suitable for regression tasks. Specifically, we use `linear regression` and `Multi-Layer Perceptron`, meta classifiers, such as `Bagging`, `Stacking` and `Random Subspace`, and decision trees, such as `M5P` and `Random Forest`. We obtain the best result and benchmark our approach with `M5P` - a model tree algorithm used for numeric prediction (Table 1). We reach MAE of **0.0638** (Pearson's score of 0.7811), outperforming the baseline model of the Shared Task (Section 3).

Next, we experiment with different feature groups and combinations with the goal to select the optimal feature subset. We train with the five best performing algorithms in each step but report only the results of the best model.

4.3.2 Ablation Studies

We narrow down the selection for the best performing features based on the three feature groups: *frequency* features, *linguistic* features and *word embeddings*.

Classifier	Pearson	MAE
M5P	**0.7811**	**0.0638**
Random SubSpace	0.77	0.0657
Bagging	0.7693	0.0657
Random Forest	0.7655	0.0661
Decision Table	0.7601	0.0665

Table 1: 10-fold cross-validation results on the training set for the top 5 classifiers

We pay special attention to frequency features since the previous work shows that word frequencies are usually among the most informative features (Yimam et al., 2018). First, to figure out the best way to represent frequencies of lower cased word forms, we train the M5P model on different frequency representations: *log relative, raw, ZIPF* and *frequency rank*. We use only the best frequency representation, log relative frequency, in the following steps. We then test the models with frequencies from various sources.

We also conduct experiments to understand the impact of word embeddings using 300-dimension pre-trained word vectors[12], and BERT[13] embeddings, where we concatenate layers 7 and 11 (Chronis and Erk, 2020) which gives better results than concatenating or summing the last four hidden layers.

We then conduct the final ablation study. Given the complete set of features, we employ WEKA's feature selection algorithms and remove the least informative features, one feature at a time. In case it does not result in an improvement, the feature is added back and we continue with the next available feature.

5 Results

In this section, we present our results and discuss the key findings. All discussed systems are trained with the `Random Forest` classifier.

5.1 Frequency Features

We find that a combination of frequency features from different sources alone can result in high performance (Table 2). In this case, daily spoken language sources, such as film subtitles, seem to be the most informative. However, adding more frequency features does not necessarily improve the results (Tables 2 and 3).

[12] fastText for English

[13] We use `bert-base-uncased` from Hugging Face (Wolf et al., 2020)

Frequency Sources	Pearson
All - EMEA	**0.713**
All	0.7128
All - EMEA - Bible	0.7041
OpenSubs + BNC + EnWiki + SimpleWiki	0.6882
OpenSubs	0.6536
SubIMDB	0.6479

Table 2: Frequency Sources

Features	Pearson
9 frequencies + corpus + POS + syllCount + charCount (*13 features*)	0.764
Above + BERT 7-11 (*1550 features*)	0.6953
9 frequencies + corpus + POS + sentence features - depRel - distToHead - NER (*44 features*)	0.7907
Above - imdbFreq	0.7909
Above - CEFR vocabulary percentages	0.7921
Above - freqPm	0.7924
Above - harmonicMeanDiff (*36 features*)	**0.7925**
Above + best BERT 7-11 (*76 features*)	**0.7942**

Table 3: Feature Ablation Experiments

5.2 Linguistic Features

During the experiments with the linguistic features, we obtain the best results using a reduced 36 feature combination (Table 4). We find out that syntactic features such as target word *distance* to the syntactic head of the sentence and its *syntactic relation* to the head of the sentence seem to worsen the performance (Table 3). The full list of ablation steps can be found in Appendix A.

Furthermore, removing the *sentence-level* features results in a slight decrease of the overall performance (from Pearson's score of 0.7925 to 0.7791). It indicates that either word-level information remains the most informative for this task or that a single sentence does not provide sufficient contextual information.

5.3 Word Embedding Features

Table 4 shows results for the best systems that are trained on *linguistic* features only, *word embedding* features only and the *combined* set of features.

The system trained on the word embeddings performs significantly worse than the other two systems. BERT embeddings only improve the result if we select a subset of 76 out of the 1536 embedding features with WEKA's `CfsSubsetEval` (Hall, 1999). The model trained on the combined set of features performs the best, reaching Pearson's score of **0.7942**. However, the difference between this system and the one trained on linguistic features is statistically insignificant. These results indicate that word embeddings are less informative than the linguistic information. Additionally, word embedding computation can be costly in terms of the added complexity and the computational resources. We, therefore, argue that a simpler feature combination is sufficient and submit our second best model to the Shared Task.

Feature Combination	#Features	Pearson
36 Linguistic + 76 Embedding	112	**0.7942**
Linguistic	36	0.7925
BERT Embeddings	1536	0.6999

Table 4: Best systems trained on linguistic, word embedding and the combined features

5.4 Test Set

The submitted system that is trained on 36 linguistic features (Appendix B) is evaluated on the official Shared Task test set and reaches Pearson's score of **0.7588**, ranking in the upper half of the submitted systems.

6 Conclusion

In this paper, we have described the design of our system submitted to the LCP Shared Task 2021 and discussed the key findings of our feature engineering approach. We aimed to design a simple system that would not require much classifier tuning or complex feature computations. Our two best models are trained on the `Random Forest` classifier with the default hyperparameters. The best system is trained on a 112 feature set which includes word embeddings. The second best system is trained on a simple 36 linguistic feature set. We submit the simple system since the performance difference between the two systems is not significant. The model is placed in the upper half of the Shared Task rankings for the single-word prediction subtask (Pearson's score of 0.7588), demonstrating how a simple approach can achieve high performance results.

Further analysis of the feature ablation studies confirms that word frequencies seem to be the most informative among all features. We also observe that even though including contextual information does improve the overall result, the performance differences are small. Future research might therefore look into including more contextual information than one sentence. In addition, the perception of word complexity differs from reader to reader. Future work could target specific reader groups, such as people with dyslexia or second language learners. In this case, the relevant background information of the readers should be included in the annotation and experimentation processes.

References

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1):1–20.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The Bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. *The WEKA Workbench*. Morgan Kaufmann. Fourth Edition.

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*, pages 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Andrew Hall. 1999. Correlation-based feature selection for machine learning.

Tomoyuki Kajiwara and Mamoru Komachi. 2018. Complex word identification based on frequency in a learner corpus. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 195–199, New Orleans, Louisiana. Association for Computational Linguistics.

George R Klare. 1974. Assessing readability. *Reading research quarterly*, pages 62–102.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior research methods*, 44(4):978–990.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Gustavo Paetzold and Lucia Specia. 2013. Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

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)*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

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*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 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.

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)*.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *International Symposium on Biological and Medical Data Analysis*, pages 184–192. Springer.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.

Appendices

A Feature Ablation Experiments

Features	#Features	Pearson	Features permanently removed
9 frequencies + corpus + POS + syllCount + charCount	13	0.764	
Above + BERT 7-11	1550	0.6953	
9 frequencies + corpus + POS + sentence-level features (- deprel - distToHead - NER)	44	0.7907	yes
above + 300 fastText word embeddings	344	0.766	
44 - imdbFreq	43	0.7909	yes
43 - Oxford lists - Kelly lists	32	0.7902	
43 - AoA	42	0.7891	
43 - CEFR vocabulary percentages	38	0.7921	yes
38 - avgSentenceLength	37	0.792	
38 - linsearWrite	37	0.7917	
38 - unknownWordPercentage	37	0.7906	
38 - difficultWordPercentage	37	0.7914	
38 - lexicalDiversityRate	37	0.792	
38 - textComplexityScore	37	0.7919	
38 - countPhones	37	0.7918	
38 - percKnown	37	0.7908	
38 - freqPm	37	0.7924	yes
37 - prevalence	36	0.79	
37 - freqZipfUS	36	0.7923	
37 - avgDiffRating	36	0.7923	
37 - harmonicMeanDiffRating	36	**0.7925**	yes
36 + best BERT 7-11 (76)	112	**0.7942**	yes

B Final Feature Set

Feature	Description
corpus	One of {bible, biomed, europarl}
POS	Part-of-speech tag
linsearWrite	readability measure used in U.S. military
avgSentenceLength	number of words in the sentence
unknownWordPercentage	unknown word percentage
difficultWordPercentage	difficult word percentage
lexicalDiversityRate	type token ratio (unique words/all words)
textComplexityScore	normalised sum of all sentence-level scores
countPhones	count of phones in word
AoA	age of acquisition
percKnown	Percentage of population that knows the word.
prevalence	word prevalence
freqZipfUS	ZIPF frequency calculated from the AoA dataset
avgDiffRating	Average of difficulty ratings from SVL 12000 dataset
kelly_a1 oxford_a1 kelly_a2 oxford_a2 kelly_b1 oxford_b1 kelly_b2 oxford_b2 kelly_c1 oxford_c1 kelly_c2	*boolean: for word that occurs in the CEFR wordlist*
syllCount	number of syllables in the word
charCount	number of characters in the word
Europarl_log_rel_freq BNC_log_rel_freq OpenSubs_log_rel_freq SimpleWiki_log_rel_freq EnWiki_log_rel_freq SubIMDB_log_rel_freq News_Comm_log_rel_freq bible_log_rel_freq	*log relative frequency of word in the corpus*
complexityTargetClass	numeric