katildakat at SemEval-2021 Task 1: Lexical Complexity Prediction of Single Words and Multi-Word Expressions in English

Katja Voskoboinik
Aalto University
Helsinki, Finland
ekaterina.voskoboinik@aalto.fi

Abstract

This paper describes systems submitted to SemEval 2021 Task 1: Lexical Complexity Prediction (LCP). We compare a linear and a non-linear regression models trained to work for both tracks of the task. We show that both systems are able to generalize better when supplied with information about complexities of single word and multi-word expression (MWE) targets simultaneously. This approach proved to be the most beneficial for multi-word expression targets. We also demonstrate that some hand-crafted features differ in their importance for the target types.

1 Introduction

SemEval-2021 Task 1 is the task of Lexical Complexity Prediction (LCP) (Shardlow et al., 2021). The goal of the task is to assign a target in a context a continuous value ranging between 0 and 1, where 1 indicates complete unintelligibility and 0 signals perfect familiarity as perceived by a native speaker. The task has two tracks: predicting the complexity score of single words and predicting the complexity score of multi-word expressions (MWE). Such a task can be useful in applications like text simplification or automatic language proficiency evaluation.

The CompLex dataset (Shardlow et al., 2020) used in this task is the first English dataset for the task of LCP. The dataset contains single words and MWEs annotated with their lexical complexity score in a specific context. The annotations were provided by native speakers of English. The targets and their contexts were obtained from texts of different domains: the Bible, Europarl, and biomedical texts. The dataset opens several avenues for research. For instance, how does the perceived complexity of single words and MWEs differ? How does context affect the lexical difficulty of a target? How does text genre affect comprehensibility of words?

We were interested if there is a difference in performance when using the same representation methods for single words and MWEs. It was decided to approach both tracks as the same problem. We did not distinguish between MWEs and single words and they both were treated as one lexical unit. The same array of features was extracted to represent the targets and linear and non-linear regressors were trained using both subcorpora. This strategy showed performance gains for both single targets and MWEs.[1]

In addition, we wanted to investigate how much the classic hand-crafted features like frequency and length together with subword information and contextualized embeddings (not employed previously for LCP) contribute to complexity estimation of both single words and MWEs. We present the analysis of feature imortance rankings in Section 6.

2 Related Work

Complex Word Identification (CWI) is the task of determining how difficult a lexical unit is to a target audience (Shardlow, 2013). The knowledge about the lexical unit complexity can benefit several NLP tasks such as text simplification (TS) or applications related to second-language (L2) acquisition.

The goal of TS is to adapt a text to make information, for example, news, more accessible for readers. The TS target group can be language learners (Petersen and Ostendorf, 2007), people with cognitive disabilities (Yaneva et al., 2016) or people with low literacy skills (Aluisio et al., 2010). One of the strategies of TS is lexical simplification (LS). LS is the task of substituting complex words with simpler ones without changing the original mean-

[1] The code and notes are available at https://github.com/katildakat/COMPLEX

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 700–705
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

ing. To perform the LS one should first identify the units that might pose a difficulty (Shardlow, 2013).

In the area of L2 acquisition, lexical complexity information can be used both for generating study materials appropriate for a learner's level (Alfter and Volodina, 2018) and for evaluating how proficient a student is (del Río, 2019).

To understand what makes a word complex for a target audience, one needs to obtain data with complexity annotated. The first manually labelled resource for CWI (CWI 2016) was introduced in SemEval-2016 Task 11 (Paetzold and Specia, 2016). It contained sentences with words marked by non-native speakers of English as either difficult or easy to understand. A word was labelled as complex if at least one annotator marked it as such. Thus, words were classified in a binary fashion without addressing the proficiency levels or native languages of the annotators.

Another CWI dataset (CWI 2018) was presented for BEA workshop 2018 (Yimam et al., 2018). It contains CWIG3G2 datasets (Yimam et al., 2017) expanded further with the French subcorpus. This is a multilingual dataset with words and MWEs marked as complex or simple within a given context. There is no standard form for MWEs. The annotators were free to label any sequence of words as a difficult MWE. The complexity judgments were collected from native and non-native speakers. In addition to the binary labels, the words were also assigned an aggregated complexity score. The score was computed as the proportion of annotators that found a word complex.

In summary, in both CWI 2016 and CWI 2018 the annotators were not asked to provide a degree of difficulty. The MWEs in CWI 2018 were not clearly defined making the nature of their complexity hard to investigate. The CompLex dataset used in SemEval-2021 Task 1 was constructed to amend the aforementioned faults of CWI 2016 and CWI 2018. First, it treats complexity as a continuous value. Second, it bounds MWEs to only pairs of adjective-noun or noun-noun phrases allowing for more targeted research. We believe that models trained using CompLex have more flexibility in their application. For example, one could set a threshold of complexity to account for different language proficiency levels for both TS and L2 acquisition-related applications.

One of the goals of CWI is to establish what makes a word complex. The reports for CWI 2016 (Paetzold and Specia, 2016) and CWI 2018 (Yimam et al., 2018) as well as the investigation of CWI 2016 results (Paetzold and Specia, 2016) show that such features as frequency and length are the most predictive for establishing word complexity. Moreover, according to the baselines provided by SemEval-2016 Task 11 organizers (Paetzold and Specia, 2016), the degree of polysemy of a word was also quite successful. In addition to hand-crafted features, the teams in both competitions made use of different static word embeddings but they didn't outperform the frequency-based features.

3 System overview

A linear and a non-linear models were compared. We have trained a linear regression and a multilayer perceptron using the same array of features. The features can be divided into two categories: embeddings and hand-crafted features for both target and context.

3.1 BERT Embeddings

For embeddings, it was decided to represent targets and their context using BERT model (Devlin et al., 2018). First, BERT is able to provide a target with a representation dependent on the context. Second, because of its next sentence prediction objective during training, it is also able to produce a separate representation for the whole sentence in the same vector space as a target. We were interested to see if target representation would benefit from combination with this additional context information.

When a single word target is present in the BERT's token vocabulary, it is represented simply as a vector assigned to it by the model. In the case when a target is absent from the BERT's vocabulary, it is represented as an average vector of its subword embeddings. MWE targets were always represented as an average representation for their BERT tokens. Contexts were assigned with [CLS] token embeddings. Finally, target and context embeddings were combined into a mean vector of 768 dimensions. When used as features to represent the training dataset in the linear regression model, mean embeddings demonstrated a slightly better performance on the trial data than concatenated vectors. For this reason, we have opted for the average embeddings of targets and contexts instead of the concatenation.

3.2 Hand-crafted features

In addition to contextualized embeddings, we were also interested to explore other features. We were especially interested to study how the target's subword information can be used to explain its complexity value. The final set of features was as follows:

1. The number of BERT vocabulary tokens (an average number for MWE) in a target. This feature was chosen because it implicitly contains frequency information about a target. BERT uses WordPiece tokenizer (Wu et al., 2016). WordPiece is a frequency-based word segmentation algorithm. It learns to unite substrings into new vocabulary items to increase the likelihood of its training data. This means the targets that were tokenized into several BERT tokens were infrequent in the tokenizer's training data.

2. A BERT score for a masked target. This feature was intended to convey information about how easy it is to predict a target in a given context. This approach however has a downside for our specific BERT implementation: BERT base model was trained to predict a randomly masked WordPiece token not a whole word token. It was decided that all single word targets are to be replaced with one MASK token. An average log probability to appear in place of a mask for every target subtoken was collected. MWE targets were substituted with two MASK tokens. An average log probability for subtokens of both words is collected, summed and divided by two.

3. A number of subwords a target is divided into (an average number for MWE) by a Morfessor segmentation model (Virpioja, 2013). This feature was expected to be a better complexity predictor than a target length in characters since it might be able to indicate a number of word parts connected to semantic or grammatical meaning.

4. An average frequency of subwords a target contains. This feature was expected to reflect how easy it would be to derive a meaning from word subparts. The frequencies were estimated using the segmentation model. In CWI 2018 the character n-gram frequency information employed by (Alfter and Pilán, 2018) achieved high results. The success of this approach might be supported by the evidence that morphological awareness affects how both native and non-native speakers process words (Kimppa et al., 2019) (Deacon et al., 2014). This feature was chosen to investigate if frequency of morpheme-like subwords is a also a good lexical complexity predictor.

5. A number of WordNet synsets (Fellbaum, 1998) that a target is present in. This feature was used to provide the information on the target's degree of polysemy. For the MWEs, we counted both synsets for the whole expression as well as synsets where either of the parts is present.

6. The length of a target in characters (an average length for MWEs).

7. Finally, we have chosen to include word frequency (an average frequency for MWEs). Frequency information is known to be a good predictor for complexity, so it was reasonable to use it as a baseline to compare other features to.

For the submitted system, the embedding features were concatenated with hand-crafted features into 775 dimensional vector. This vector was used as an input to both regressor types.

We have also investigated if the described setup would benefit from feature selection. We decided to half the original feature vector's size in half by leaving only 400 most informative dimensions. For the linear model, the dimensions were ranked by their F-score. The mutual information was used to chose the dimensions for the neural model.

4 Experimental setup

4.1 Data

During the system development phase, models were trained only with the train subsets of the data, and then their performance was evaluated on trial subsets. For the final submission, both linear and non-linear models were trained with all the data available (single target train and trial, MWE target train and trial).

	LR all	NN all	LR 400	NN 400
Singles	0.666	**0.712**	0.688	0.707
MWEs	0.783	0.785	**0.796**	0.774

Table 1: Joint System Results

	LR	NN
Singles	0.669	**0.706**
MWEs	0.678	**0.752**

Table 2: Separate Systems Results

4.2 Parameters and Tools

Both linear regression and neural network models were trained with scikit-learn 0.24.0 [2]. It was also used for the feature selection process. The neural network model is a simple Multilayer Perceptron regressor with one ReLu layer of 20 neurons and alpha parameter set to 0.9. Using 8-fold cross validation procedure on all labelled data, we noticed that smaller layer sizes and larger alpha parameters produced better results. However, we feel it is important to note that the hyperparameters were not tuned with proper care, and we believe that better configurations might be possible.

We used BERT base model (cased) to get contextualized embeddings. The cased model was chosen because in CompLex dataset case plays an important role when distinguishing a target from other words in context. The model was used through the 4.0.0 version of transformers library (Wolf et al., 2020).

The Morfessor segmentation model was trained with Morfessor 2.0[3] using logarithmic frequency dampening for words in the data, and with the corpus weight parameter $\alpha = 0.1$. The text used for the segmentation model comes from samples of all subcorpora in The Corpus of Contemporary American English (COCA) but for the Academic texts[4].

The WordNet was used via NLTK 3.4[5]. The word frequencies were obtained using the Zipf frequency estimates in the 'best' wordlist of wordfreq library[6] (Speer et al., 2018).

5 Results

The results of the systems trained jointly with single and MWE targets are presented in Table 1. The results for the systems trained with each subcorpus separately are reported in Table 2. The results in both tables are given using Pearson correlation coefficients for the test data indicated by row names. LR stands for the linear regression model, and NN stands for the neural regressor. Captions 'all' and '400' distinguish between models trained using all 775 dimensions or 400 with the best scores.

	S+MWE		S		MWE	
FEATURE	m	f	m	f	m	f
len_tok	1	1	1	1	1	1
bert_prob	2	2	2	2	3	2
morf_len	3	3	4	3	2	3
morf_freq	9	89	8	53	109	338
n_senses	5	13	3	4	5	5
len_char	4	110	5	197	8	11
word_freq	0	0	0	0	0	0

Table 3: Importance Ranks for Hand-crafted Features

For the single word track, the CodaLab system for some reason accepted only the linear scores, and for the MWE track CodaLab, conversely, accepted only non-linear model predictions. Moreover, the top score for the linear model in the single word track was reported without using Morfessor features.

6 Discussion

As can be seen from the results tables, two trends are obvious: MWE targets always benefit from being trained together with single word targets, and the non-linear model tends to slightly outperform the linear one. This can be contributed both to the nature of MWE and to the smaller size of the MWE subcorpus. Although the linear system trained with only single word targets showed better results than the joint one, the non-linear model has also gained from the information about both types of targets when predicting single word complexity. Finally, the feature selection procedure was able to improve the performance of the linear model.

We were interested to find what features for MWEs and single words signal lexical complexity in a similar manner and what features differ in their usefulness. For this purpose, we collected F-scores and mutual information values for the hand-crafted features evaluated for the joint dataset as well as

[2] scikit-learn https://scikit-learn.org/stable/index.html

[3] Morfessor 2.0 morfessor.readthedocs.io

[4] COCA samples www.corpusdata.org/formats.asp

[5] NLTK https://www.nltk.org/

[6] wordfreq https://pypi.org/project/wordfreq/

for the target type subcorpora.

The ranks of hand-crafted features according to their F-scores and mutual information values can be found in Table 3. The features are listed in the same order as they were presented in Section 3.2. The names of the columns reflect the content of the dataset divisions that were explored: S stands for the single targets part of the data, MWE stands for the examples with only MWE targets, and S+MWE marks the results for the joint dataset. The mutual information ranks can be found in 'm' columns and F-scores are given in 'f' columns. The ranks are reported for the features evaluated with the train and the trial parts of the corpus simultaneously.

The information collected in Table 3 shows some differences in feature importance for predicting single word targets and MWE targets, as well as, differences in how suitable some features are for linear and non-linear models. Moreover, the consistency of how high most of the hand-crafted features rank indicates that they remain relevant for LCP and CWI tasks even in presence of such modern approaches as contextualized embeddings.

All hand-crafted features were present in the top 20 highest scored dimensions with mutual information for the joint data and for the single target data. For the MWE targets, information about morph frequency played a less important role placed at only 110 place. Moreover, with the F-score rankings, morph frequency was absent in the top 20 dimensions from all the data configurations.

Target length in characters has not appeared in the top 20 most correlated features for the joint dataset and for the single targets, but it was still relevant for MWEs. Word frequency was rated as the highest correlated feature in all setups, it was followed by the BERT token number feature. These two features were followed by morph number and by the probability of a masked target. Surprisingly, the subword frequency feature was not as successful. The reason for this can be the small amount of data it was estimated on.

7 Conclusion

This paper presents the results of two systems submitted to SemEval 2021 Task 1: Lexical Complexity Prediction (LCP). We show that training a system jointly with single word targets and MWE targets benefits the predictability of both target types, especially the MWEs. We also show that the frequency of subwords feature is more predictive for single targets, while length of a target in characters is more useful for MWE complexity estimation. Finally, we show that classic frequency feature is still the most predictive one, even when used together with new contextualized embeddings.

For the future work, we would like to explore if the underwhelming results of the subword frequency feature can be amended by collecting statistics from a larger resource. Another thing we would like to research is what makes the joint training with single and MWE targets successful. Is it the smaller amount of data available for MWEs? Or is it the nature of noun-noun and adjective-noun MWE expressions? Does the second word of the pair contribute more to the MWE complexity and thus compares better to single word targets?

8 Acknowledgments

I would like to express my gratitude to the anonymous reviewer for their insightful and helpful comments. I also want thank Ragheb Al-Ghezi for proofreading and Mikko Kurimo for his comments on an earlier draft of the paper. Lastly, I want to say thank you to Science-IT team for providing computational environment used in the experiments. This work is a part of DigiTala project which is funded by the Academy of Finland (grant number 322625).

References

David Alfter and Ildikó Pilán. 2018. SB@GU at the complex word identification 2018 shared task. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 315–321, New Orleans, Louisiana. Association for Computational Linguistics.

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88, New Orleans, Louisiana. Association for Computational Linguistics.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Hélène Deacon, Michael Kieffer, and Annie Laroche. 2014. The relation between morphological awareness and reading comprehension: Evidence from mediation and longitudinal models. *Scientific Studies of Reading*, 18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Lilli Kimppa, Yury Shtyrov, Suzanne Hut, Laura Hedlund, Miika Leminen, and Alina Leminen. 2019. Acquisition of l2 morphology by adult language learners. *Cortex*, 116.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *In Proceedings of Workshop on Speech and Language Technology for Education*.

Iria del Río. 2019. Linguistic features and proficiency classification in L2 Spanish and L2Portuguese. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 31–40, Turku, Finland. LiU Electronic Press.

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. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*.

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)*.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. Luminosoinsight/wordfreq: v2.2.

Peter; Grönroos Stig-Arne; Kurimo Mikko Virpioja, Sami; Smit. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. D4 julkaistu kehittämis- tai tutkimusraportti tai -selvitys.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 293–299, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 - complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.