archer at SemEval-2021 Task 1: Contextualising Lexical Complexity

Irene Russo
ILC-CNR, Pisa, Italy
irene.russo@ilc.cnr.it

Abstract

Evaluating the complexity of a target word in a sentential context is the aim of the Lexical Complexity Prediction task at SemEval-2021. This paper presents the system created to assess single words lexical complexity, combining linguistic and psycholinguistic variables in a set of experiments involving random forest and XGboost regressors.
Beyond encoding out-of-context information about the lemma, we implemented features based on pre-trained language models to model the target word's in-context complexity.

1 Introduction

Lexical complexity prediction is the task aiming at evaluating the complexity of a word in context, modeling a crucial aspect of reading comprehension. Complex words can slow down the reading process; there is a well-known correlation between a word's difficulty and the time spent looking at it, as emerging from eye-tracking experiments (Mousikou et al., 2021).

Assessing the complexity of a specific word in context is a crucial prerequisite for NLP systems aiming to evaluate a text's readability and produce a simplified version of it. It is a prerequisite for text simplification systems based on lexical substitutions and can be the starting point to tailor a text to the user's needs. It is a topic worthy of investigation from multiple points of view. In the past, datasets containing crowdsourced evaluations of lexical items' complexity (Paetzold and Specia, 2016b; Štajner et al., 2018) have been used in evaluation campaigns.

In this paper, we introduce the system used to assess single English words lexical complexity at SemEval-2021 Lexical Complexity Prediction task (Shardlow et al., 2021). Based on previous approaches to this issue, we combine linguistic and psycholinguistic variables, using a random forest regressor and an XGboost regressor. The majority of the variables encode information about the word out-of-context (e.g., frequency, number of letters, age of acquisition) without considering the sentential context and how it can affect an item's complexity.

To approximate in-context complexity, we consider the cloze probability of a word as its probability to complete a particular sentence frame. We experiment with different language models in a masked word prediction framework, taking into account the first ten most probable words occurring in that context.

2 Related works

A wide range of approaches has been used for lexical complexity prediction in past evaluation campaigns. However, previous tasks focused on classification since the proposed datasets labeled words in context as easy or difficult.

Including more classes makes the task more difficult. (Garí Soler et al., 2018) investigate the role of word embeddings in lexical complexity prediction for French words, using as training sets two French lexical resources that encode the distribution of words across different levels of difficulty. According to the authors, the task is influenced by the context of use of the words. Word embeddings, encoding contextual information, can be helpful to determine the lexical complexity of target words. The authors experimented with different neural network settings (with and without hidden layers), using as features the number of characters, the number of phonemes, and the log frequency in a corpus of film subtitles plus word embeddings trained on Wikipedia with fastText. However, the combination of word embeddings with such features does not improve the results compared with other sets

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 694–699
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

of features that always contain frequency, a crucial indicator of lexical complexity.

In past evaluation campaigns, there have been occasional attempts at incorporating word embeddings models in the automatic evaluation process, with the assumption that lexical complexity should be evaluated as a contextual variable. However, better results have been obtained considering just static, out-of-context properties of lemmas.

More recently, the trend to use embeddings from language models to predict psycholinguistic variables can provide insights about how to incorporate them in experiments aiming at understanding the complexity of human comprehension (Hao et al., 2020).

However, if we frame lexical complexity as a measure strongly dependent on words' psycholinguistic properties, we should recognize that past computational efforts for predicting word norms did not take into account the role of context (Russo, 2020; Charbonnier and Wartena, 2019). Static word embeddings such as word2vec have been used to predict values of psycholinguist norms usually assessed in experimental settings (Ljubešić et al., 2018; Rothe and Schütze, 2016). More recent Transformed-based language models that consistently incorporate contextual knowledge have not yet been considered for this task.

3 The Dataset

The Lexical Complexity Prediction shared task at SemEval-2021 (LCP-2021) (Shardlow et al., 2021) is based on an English dataset with a 5-point Likert scale annotation. The complexity score is similar to that included in another dataset (Shardlow et al., 2020). It ranges from very easy for very familiar words to very difficult (unclear words that an annotator had never seen before). Annotators are explicitly invited to evaluate the role of the sentence in inferring the meaning of the word. The task is structured into two sub-tasks:

- Sub-task 1: predicting the complexity score of single words;
- Sub-task 2: predicting the complexity score of multi-word expressions.

The task is inspired by two previous competitions (CWI 2016 and CWI 2018) about boolean complex word identification, aiming at identifying which words are likely to be considered complex or not by a given target population. However, in LCP-2021 lexical complexity is a continuous property, and the task consists of predicting the complexity score for each target word in context.

domain	mean	std dev
bible	0.296	0.132
europarl	0.287	0.109
biomed	0.325	0.152

Table 1: Mean complexity and standard deviation for each domain in the LCP-2021 training dataset.

7,662 sentences and 3,298 unique tokens compose the LCP-2021 training dataset for single words evaluation: each token appears in more than one sentence, making the impact of context crucial especially for subsets of sentences with highly variable values. For example, the word *livers* occurs 2 times in the dataset, with different complexity scores:

- *The activity of BCKDH in* ***livers*** *of homozygous knockout mouse pups was undetectable, accounting for the accumulation of unmetabolized BCAA.* (complexity score = 0.0499)
- *The comparison of gene expression in* ***livers*** *of mock- or cadmium-treated Mtf1Mx-cre and Mtf1loxP mice revealed several MTF-1 target gene candidates.* (complexity score = 0.323)

Sentences are extracted from three domains: the Bible, the English part of the European Parliament proceedings, and a biomedical corpus composed of scientific papers. Table 1 reports the mean complexity and the standard deviation for each domain. Target words extracted from the biomedical corpus are the most complex. Due to the variability in complexity for the same target word, the biomedical corpus is also the domain that poses significant challenges.

We propose a system for sub-task 1, encoding for each target word numerical values concerning a set of variables described in Section 4. We do not propose a system for sub-task 2.

4 Out-of-Context and In-Context Lexical Complexity

The lexical complexity of a word can be represented as an out-of-context property of a lemma or an in-context property of a word.

Following the first approach, the same lemma has a fixed lexical complexity value, depending on features such as the number of characters or senses in WordNet (see the list of out-of-context features below).

When considering the word in context, its disambiguation could affect its complexity rating because a sense could be more complex than the others (for example, when a word is used in its specialized sense). However, because of the lack of methodologies for assessing senses' complexity and the unsatisfactory performance of word sense disambiguation systems, the role of senses' complexity for lexical complexity prediction can not be investigated. Several systems participating at CWI2018 took into account the role of context, focusing on the whole sentence where target items occur. However, quite interestingly, one of the best models (Gooding and Kochmar, 2018) does not consider the influence of the textual context for determining the target word's complexity. We provide the list of out-of-context features used in our system:

- Length: length of each target word (number of characters);
- Syllables: number of syllables of each target word[1];
- length sentence: length of each sentence (number of tokens);
- Word freq: frequency of the target word in the Exquisite Corpus[2];
- AoA Kup: age of acquisition (AoA) of the target word in (Kuperman et al., 2012) dataset. The age of acquisition of a word is a psycholinguistic variable concerning the age at which a word is typically learned. We assume that easy words are learned at a younger age;
- Children freq: the natural logarithm of the frequency of lemmas in children movies subtitles included in a corpus of subtitles(Paetzold and Specia, 2016a). We expect that difficult words will be less frequent in this corpus;
- Visual Genome (VG) freq: the natural logarithm of the lemmas' frequency in the Visual Genome descriptions corpus. The Visual Genome dataset (Krishna et al., 2017) is the largest dataset of image descriptions for English. It is composed of dense annotations of objects, attributes, and relationships between objects for 108K images. As a preprocessing step, the descriptions have been annotated with TreeTagger (Schmid, 1994) and the list of lemmas has been ordered by frequency;
- ImageNet: the presence of the target word in ImageNet (boolean feature). Only concrete nouns can be included as pictures in this resource. We assume that easy words tend to be more frequently concrete (Russakovsky et al., 2015);
- Uppercase: this variable takes into account the relative number of uppercase letters in the target words, and it is used to detect acronyms. We notice that acronyms are generally rated as difficult word;
- Scrabble: for each target word, its value according to Scrabble's rules. In this word game, each letter's number of points is based on the letter's frequency in standard English. We expect that complex words will have higher ratings;
- Senses: number of senses of the target word in WordNet (Fellbaum, 1998);
- Bible/europarl/biomed: boolean feature encoding if the sentence belongs to one of these domains.

A word x is difficult in a sentential context if the reader has never encountered it. The relative frequency should be sufficient to explain the perceived complexity of x). A word can also be difficult if its meaning in that specific context is not the most common one, i.e when a specialized sense is accessed or a metaphorical meaning is created. If we know a word by the company it keeps, we do not know a word when its company is somewhat eccentric.

Lexical complexity as an in-context property can be modeled considering the influence of the surrounding text. There are two ways to model the textual context's influence on the lexical complexity of a target word: local context (a window span surrounding the target word) and global context (the whole sentence). In the first case, words surrounding the target word can increase the overall

[1] The values are obtained using syllables 0.1.0 https://pypi.org/project/syllables/

[2] The values are obtained using wordfreq 2.3.2 https://pypi.org/project/wordfreq/

complexity in that text span. In the second case, the probability of a word in a masked word prediction task that concerns the whole sentence can be a good approximation of the intuition that words semantically difficult to generate are more complex than the simpler ones. We implemented two types of variables as in-context features to address in-context lexical complexity:

- Position [Language Model]: The target word's position among the first ten most probable words completing the sentence in a masked context for five language models. The language models tested are BERT, XLNet large, BART, ELECTRA, and RoBERTa. We used pre-trained models made available by HuggingFace;
- Out-of-context complexity of the previous tokens: the value is obtained by selecting the five content words (nouns, adjectives, or verbs) preceding the target word and averaging their complexity values resulting from a random forest regressor that includes just out-of-context variables.

The Pearson correlations among each feature and target word's lexical complexity reveal that word frequencies are the most relevant features, especially frequencies extracted from children movies' subtitles (see Figure 1). The age of acquisition of words is another variable strongly correlated with the complexity of the target words (r=0.55).

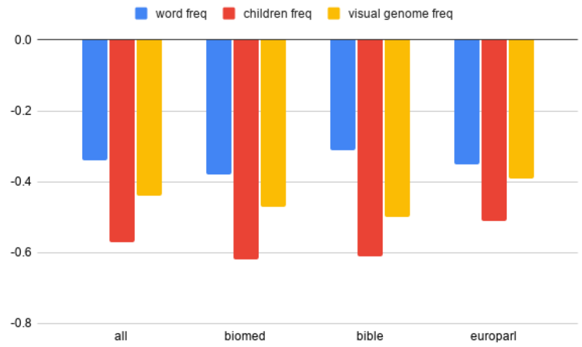

Figure 1: Pearson correlations between lexical complexity and word frequencies from different corpora, reported for each LCP-2021 domain.

5 Experiments

The set of features described in Section 4 has been implemented for the training set, tested on trial set, and - to avoid overfitting - on multiple training-test splits with test sets mimicking the trial set's composition. We experimented with a system based on a random forest regressor (RF), and a system based on an XGboost regressor (Chen and Guestrin, 2016), both implemented in sklearn. For the RF regressor, we choose to measure the quality of a split with mean absolute error. We also normalized the features with the standard scaler function (scaling each feature between 0 and 1). We obtained comparable results, with random forest regressor performing slightly better for several training-trail splits. For this reason, Table 2 summarises RF results. We report the mean absolute error (MAE) and Pearson correlation (R), used to rank the systems at LCP-2021 task.

Domain	MAE	R
bible	0.075	0.69
europarl	0.054	0.76
biomed	0.066	0.86
all_out_of_context	0.063	0.817
all_in_context	0.095	0.415
all	0.065	0.80

Table 2: Random forest regression results on trial set.

features	MAE	R
all_out_of_context	0.063	0.793
all	0.062	0.799

Table 3: Average random forest regression results on five training-trial splits.

In-context features emerge as useless from these results; however, testing with different trial sets, we infer that this set of features could improve the performance (see Table 3) and, as a consequence, we included all the features for the processing of the test set released by LCP-2021 organisers. Concerning the role of word frequencies, that are negatively correlated with lexical complexity (see Section 4), frequencies from a general corpus used together with frequencies from children movies subtitles guarantee a good performance of the RF regressor in terms of MAE and Pearson correlation (see Table 4). Our system ranked 22 out of 54 for the single word complexity prediction task. The best result on the test set was obtained using all the features and the random forest regressor (see Table 5).

features	MAE	R
Word freq + children freq	0.069	0.74
Word freq + VG freq	0.068	0.752

Table 4: Average random forest regression results on five training-trial splits for frequency features.

all features	
Pearson	0.7561
MAE	0.0641
Spearman	0.7067
MSE	0.0069
R2	0.5707

Table 5: Best random forest regression results on test set (official results).

6 Conclusions

This paper briefly reports the system created to predict single words' complexity score for the Lexical Complexity Prediction shared task at SemEval-2021 (LCP-2021).

Our system ranked 22 out of 54 for this sub-task, with slightly inferior results to the ones obtained on trial sets. The significative role of frequencies extracted from different corpora paves the way to further investigations in this direction.

Encoding in-context complexity as a variable related to pre-trained language models' predictions had no significant impact on the results. However, in-context complexity could be modeled in different ways. Experimenting with how in-context target word's complexity changes depending on the frequencies of the surrounding words is a future analysis topic.

References

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187, Gothenburg, Sweden. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2018. A comparative study of word embeddings and other features for lexical complexity detection in French. In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 499–508, Rennes, France. ATALA.

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.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86, Online. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

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

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia. Association for Computational Linguistics.

Petroula Mousikou, Lorena Nüesch, Jana Hasenäcker, and Sascha Schroeder. 2021. Reading morphologically complex words in german: the case of particle and prefixed verbs. *Language, Cognition and Neuroscience*, 36(2):255–268.

Gustavo Paetzold and Lucia Specia. 2016a. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.

Gustavo Paetzold and Lucia Specia. 2016b. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

Sascha Rothe and Hinrich Schütze. 2016. Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the*

Association for Computational Linguistics (Volume 2: Short Papers), pages 512–517, Berlin, Germany. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Irene Russo. 2020. Guessing the age of acquisition of Italian lemmas through linear regression. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 43–48, Online. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

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)*.

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*.