An Evaluation of Binary Comparative Lexical Complexity Models

Kai North[1], **Marcos Zampieri**[1], **Matthew Shardlow**[2]
[1]Rochester Institute of Technology, Rochester, NY, USA
[2]Manchester Metropolitan University, Manchester, UK
kn1473@rit.edu

Abstract

Identifying complex words in texts is an important first step in text simplification (TS) systems. In this paper, we investigate the performance of binary comparative Lexical Complexity Prediction (LCP) models applied to a popular benchmark dataset — the CompLex 2.0 dataset used in SemEval-2021 Task 1. With the data from CompLex 2.0, we create a new dataset contain 1,940 sentences referred to as CompLex-BC. Using CompLex-BC, we train multiple models to differentiate which of two target words is more or less complex in the same sentence. A linear SVM model achieved the best performance in our experiments with an F1-score of 0.86.

1 Introduction

Children, second language learners, or individuals suffering from a reading disability, such as dyslexia or aphasia, can find certain words hard to read, interpret, or learn (Devlin and Tait, 1998; Carroll et al., 1998; Kajiwara et al., 2013; Rello et al., 2013; Malmasi et al., 2016). In readability and text simplification (TS) literature, these words are known as complex words.

Complex and non-complex words are distinguishable. Statistical, morphological, and psycholinguistic features are indicative of lexical complexity (Shardlow et al., 2022; Desai et al., 2021). Complex words are on average longer, morphologically more unique, and less frequent in general corpora, than non-complex words (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021).

With the growing popularity of distance learning platforms, the demand for new technologies that make texts more accessible for independent and remote learning has seen an exponential increase (Morris et al., 2020). Among these technologies, are TS systems that automatically simplify texts for various target populations. The first step in TS is generally referred to as lexical complexity prediction (LCP). LCP aims to identify which words in a text are complex and therefore are in need of simplification. It has been modeled as a binary classification task (Paetzold and Specia, 2016), as a regression task (Yimam et al., 2018), and more recently as a multi-class classification task (Shardlow et al., 2020).

In this paper, we explore binary comparative LCP where the goal is to determine when one target word is more or less complex than another. This new type of LCP is motivated by the need for comparative prediction methods that allow for the pairwise ranking of target words based on newly available data assigned with continuous complexity values (Shardlow et al., 2020). Binary comparative LCP aims to aid TS by improving the selection and ranking of substitute candidates for a particular complex word. It achieves this by allowing for more data to be generated from a finite dataset. For instance, a dataset consisting of 10,000 complex words assigned with complexity values can be converted into 100 million comparative instances, since every complex word can be compared to every other complex word. A binary comparative LCP classifier trained on this dataset can then make binary comparative judgements as part of a sorting algorithm. This lets a lexical simplification (LS) system effectively find the most appropriate simplification for any given complex word improving the efficiency of TS and other down-stream applications.

We aim to determine whether binary comparative LCP is possible when attempting to differentiate the complexities of two different target words in the same sentence (Table 3). We have accomplished this by adapting a recent baseline LCP dataset: CompLex 2.0 (Shardlow et al., 2020), and by asking the following research question: can LCP be modeled as a binary comparative classification task?

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 197 - 203
July 15, 2022 ©2022 Association for Computational Linguistics

The main contributions of this paper are:

1. CompLex-BC, the first binary comparative LCP dataset built from continuous data obtained through 5-point likert-scale annotation.

2. An evaluation of SVM, BERT, and BERT + MLP models for binary comparative LCP.

2 Related Work

Traditionally, LCP comes in two forms, it is either: a). a binary classification task, known as complex word identification (CWI) (Paetzold and Specia, 2016; Zampieri et al., 2017; Yimam et al., 2018), or b). a linear regression based task, simply referred to as LCP (Shardlow et al., 2021). Both CWI and LCP datasets contain target words labeled with a complexity value. This complexity value is used by a machine learning (ML) model to determine the complexity of a target word. CWI assigns a binary complexity value of either 1 (complex), or 0 (non-complex). LCP alternatively assigns a complexity value on a continuum, ranging from 0 to 1. This continuum contains multiple labels with differing complexity thresholds: very easy (0), easy (0.25), neutral (0.5), difficult (0.75), to very difficult (1) (Shardlow et al., 2020). An example is shown in Table 1.

	Folly	is	**set**	in	**great**	**dignity**
BC	1	is	0	in	0	0
CC	0.57	is	0.18	in	0.15	0.42

Table 1: Example of a sentence annotated with both binary complexity (BC) and continuous complexity (CC) values from CWI and LCP systems respectively. Target words are in bold.

Other approaches have attempted to model LCP as a multi-class classification task. Pintard and François (2020) assigned six readability levels, belonging to the Common European Framework of Reference for languages (CEFR), to target words as a means of rating their complexity for second language learners. Alfter (2021) trained a variety of models, including a convolutional neural network (CNN) and a recurrent convolutional neural network (RCNN), to predict the correct CEFR labels of target words taken from a multitude of CERF vocabulary lists.

LCP research has also included the ranking of complex words (Paetzold and Specia, 2017; Maddela and Xu, 2018). Neural regressors have been trained to identify which of two target words is more or less complex by predicting a continuous positive or negative value belonging to an inputted word pair. A positive value indicates that the first target word is more complex than the second, whereas a negative value dictates that the second target word is more complex than the first. The magnitude of the returned value also represents the degree of difference between the two target words. Positive and negative values closer to +1 or -1 respectively, show that the difference between the target words' complexities is more extreme compared to those values closer to 0 (Table 2).

Word Pair	**CC Values**	**Label**
{set, great}	{0.18, 0.15}	+0.03
{set, dignity}	{0.18, 0.42}	-0.24

Table 2: Example of two word pairs annotated with continuous comparative complexity labels. The first word pair have a similar level of complexity, whereas the second word pair have a greater disparity between their complexities.

Binary comparative LCP provides a binary complexity label that defines when the first target word is more complex (1) or less complex (0) than a second target word, be it either the same or a different target word in the same or a variety of contexts. Examples of two different target words in the same context, in this case sentence, are shown in Table 3.

Binary comparative LCP is now only recently possible due to the release of the Complex 2.0 dataset that provides a more fine-grained representation of target word complexity (Shardlow et al., 2020). This is since CompLex 2.0 is the first of its kind to contain continuous complexity values obtained through the use of a 5-point rather than a 6-point likert-scale annotation scheme that does not account for neutral labeling (Maddela and Xu, 2018), or through the use of binary annotation. Binary comparative LCP is thus a new form of complexity prediction that differs from the complexity ranking previously attempted by Paetzold and Specia (2017) and Maddela and Xu (2018), as it uses new and more fine-grained continuous complexity values to make binary comparative predictions.

Target Word 1	Target Word 2	Context	L
wood	**hyssop**	...he shall take it, and the cedar **wood**, and the scarlet, and the **hyssop**...	0
sequencing	**fly**	...the **sequencing** projects of human, mouse, rat, fruit **fly** and...	1
fish	**invertebrates**	...such as mammals, **fish**, and amphibians, but not in **invertebrates**...	0
Nehemiah	**district**	...**Nehemiah** the son of Azbuk, the ruler of half the **district** of Beth...	1
example	**avoidance**	...for **example**, a QTL for PROP **avoidance** has been suggested on...	0

Table 3: Example of the Complex-BC dataset. Target words are in bold. Only snapshots of context are shown. Label (L) 0 refers to when target word 1's complexity < target word 2's complexity, and label 1 refers to when target word 1's complexity > target word 2's complexity.

No.	Input Type	Encoding Strategies
a	Target Word only	<CLS>set<SEP>dignity<SEP>
b	Single Context	<CLS>Folly is <B>**set**<E> in great <B>**dignity**<E><SEP>
c	Two Contexts - TW	<CLS>Folly is in great dignity<SEP>Folly is set in great<SEP>

Table 4: Examples of input types and encoding strategies used.

3 Data

CompLex 2.0 The CompLex 2.0 dataset contains 9,000 instances of individual words in context. Each of its extracts were taken from the Bible (Christodouloupoulos and Steedman, 2015), biomedical articles (Koehn, 2005), and EuroParl (Bada et al., 2012). Its annotators were crowd-sourced from "the UK, USA, and Australia" (Shardlow et al., 2020).

CompLex-BC We created a new dataset containing binary comparative labels (Table 3). Complex-BC consists of 1,940 sentences that house two differing target words identified as being complex within the CompLex 2.0 dataset and that also belonged to the same sentence. Each entry comprises of a target word, a second target word, and a label. For example, given the sentence "*he shall take it, and the cedar wood, and the scarlet, and the hyssop*" from the CompLex 2.0 dataset, our new dataset adapts this sentence and provides "*wood*" as target word 1, "*hyssop*" as target word 2, and a new binary comparative label of "0" that indicates that in this sentence, target word 1: "*wood*", was rated as being less complex than target word 2: "*hyssop*" by the annotators of the CompLex 2.0 dataset (Table 3).

4 Models

We trained a SVM model given its high performance at binary CWI (Zampieri et al., 2016; Choubey and Pateria, 2016; Sanjay et al., 2016; Kuru, 2016), a BERT model (Devlin et al., 2019) per its competitive performance at LCP-2021 (Shardlow et al., 2021; Yaseen et al., 2021; Pan et al., 2021; Rao et al., 2021), and a BERT + multi-layer perceptron (MLP) model (Gu and Budhkar, 2021) to take full advantage of BERT inferred contextual features as well as the word-level features fed into our SVM model. Two naive baseline models were used to evaluate the performances of our SVM, BERT, and BERT + MLP models: a random classifier (RC) and a majority classifier (MC).

We used an integrated Intel UHD Graphics 620 GPU to train each model. Our SVM and BERT models were trained over 5 epochs. The train and test split of our dataset was set to 70:30% respectively. No target words were shared between the train and test sets.

SVM Our SVM model used a Radial Basis Function (RBF) kernel and was trained on a set of well established statistical and psycholinguistic features for LCP as these have been previously found to achieve the best results (Shardlow et al., 2022; Desai et al., 2021). These features were word length, word frequency, syllable count, average age of acquisition (AoA), and prevalence (familiarity). Word frequency being calculated in accordance to the target word's frequency in the British National Corpus (BNC) (Consortium, 2007), average AoA being calculated by averaging the AoAs within an updated version of the Living Word Vocabulary Dataset (Dale and O'Rourke, 1981; Brysbaert and Biemiller, 2017), and prevalence being calculated in accordance to the percentage of people who knew the target word as shown in the dataset provided by Brysbaert et al. (2019). These features were obtained in regards to the target word only and were not applied to any of the target word's neighbouring words.

BERT After experimenting with different hyperparameters, our BERT model (bert-base-uncased) was set to have a softmax activation layer, a batch size of 200, and a learning rate of 1e-5. Several inputs were also experimented with that took into consideration the target word along with varying degrees of contextual information. Encoding strategies adopted by the leading systems of LCP–2021 (Rao et al., 2021; Shardlow et al., 2021) and suggested by Hettiarachchi and Ranasinghe (2021), were then applied to these inputs and fed into our model.

We encoded each input into sub-word units, otherwise known as WordPiece tokens (Devlin et al., 2019). We then used the class identifier special token: <CLS>, the separator special token: <SEP>, and two custom special tokens: <B> and <E>, to distinguish between two differing target words. We referred to these as (a) the target word only, (b) single context, and (c) two contexts - TW encoding strategies respectively (Table 4). Encoding strategies (a) and (b) included the target word.

BERT + MLP As BERT assumes full sentences to encode the correct contextual information, we use a second BERT-based architecture for a fairer evaluation with the SVM model. We built a BERT + MLP model by feeding encoding strategies (a), (b), and (c) into our BERT model and then concatenate the outputted contextual features with those features used by our SVM model per Gu and Budhkar (2021). Our final BERT + MLP model then utilizes both set of contextual and word-level features for binary comparative LCP.

	Weighted Average			
Model	P	R	F1	A
SVM	**0.85**	**0.86**	**0.86**	**0.86**
BERT + MLP	0.74	0.88	0.80	0.81
BERT	0.49	0.48	0.44	0.48
MC	0.30	0.55	0.39	0.55
RC	0.50	0.50	0.50	0.50

Table 5: Best performances ranked in order of highest to least accuracy and split between test and baseline models.

5 Results

Table 5 shows the best performances of our SVM, BERT, and BERT + MLP models. Performances were measured in terms of weighted average precision (P), recall (R), and F1-score (F1). Our models' accuracies (A) were also reported.

Our SVM model achieved a F1-score of 0.86, whereas our BERT model attained a noticeably worst score of 0.44. Both models also produced drastically different accuracies of 0.86 for our SVM model and 0.48 for our BERT model with our BERT model attaining an accuracy on par with our naive MC and RC baseline classifiers. This suggests that there was not enough information for our BERT model to converge and therefore its final output labels were likely chosen at random. These performances were achieved by taking into consideration only the target words (a) and not any of their surrounding contexts (Table 4). Other encoding strategies (b) and (c) failed to surpass these performances for our BERT model.

Our BERT + MLP model achieved a greater performance compared to that attained by our standalone BERT model. It was found that by concatenating the contextual features outputted by our BERT model from single context input (b) with those features used for our SVM model, our BERT + MLP model achieved a precision, recall, f1-score, and accuracy of 0.74, 0.88, 0.80 and 0.81 respectively. Additional encoding strategies (a) and (c) attained lower performances, with an F1-score of 0.56 being returned by the former and 0.64 being achieved by the latter.

6 Analysis and Discussion

Comparing the complexities of two target words in the same sentence allowed us to fully utilize the CompLex 2.0 dataset that contained multiple instances of target words in the same context. It allowed us to generate binary comparative predictions that can be used to aid substitute selection and ranking (Sections 1 and 3).

Binary comparative LCP also allowed us to identify which words within a sentence contributed the most or the least to a context's overall complexity. Therefore, by comparing the complexities of two target words in the same sentence, we were able to identify which part of a sentence is in need of priority simplification. However, modeling binary comparative LCP in this way is not without its challenges. This is reflected in our models' performance.

Two factors are responsible for the superior performance of our SVM model in comparison to our standalone BERT model: 1). contextual similarity between the two target words, and 2). small dataset size.

One of the main advantages of transformer-based models, such as BERT, is their ability to infer bi-directional contextual relationships between a target word and its surrounding words and then use this contextual information to make accurate predictions (Devlin et al., 2019). However, since the two target words we are trying to compare are in the same sentence, BERT's inferred contextual features in relation to target word 1 were extremely similar to those inferred for target word 2, when given encoding strategy (c) (Table 4). This confused our BERT model, by making classes 0 (target word 1 is less complex) and 1 (target word 1 is more complex) hard to distinguish.

Our SVM model's word-level features of word length, word frequency, syllable count, average age of acquisition, and prevalence (familiarity) alternatively do not utilize contextual information. Instead, they rely purely on the characteristics of each target word and therefore resulted in feature representations that were more dissimilar. This had the effect of making classes 0 and 1 easier to differentiate for our SVM model and thus explaining its superior performance.

The superiority of word-level features is further demonstrated by the performance of our BERT + MLP model. Encoding strategy (b) returns from our BERT model contextual information related to the shared context (Table 4). Encoding strategy (c), however, attempts to encode contextual information belonging to each target word minus the target word.

Our BERT + MLP model performed poorly on encoding strategy (c) as it was presented from BERT two inferred contextual feature representations which were near identical. Encoding strategy (c) alternatively supplied only one inferred contextual feature representation from BERT, which allowed our BERT + MLP model to rely more heavily on its engineered word-level features. Nevertheless, encoding strategy (b) still failed to surpass our BERT + MLP model's performance beyond that of our SVM model. This indicates that the utilization of BERT's inferred contextual feature representation from encoding strategy (b), is still inferior to an SVM model using word-level features.

Another explanation for our models' performance is our dataset's size. The CompLex-BC dataset contains 1,940 instances with binary labels. Transformer-based models require large amounts of data to infer meaningful feature representations (Devlin et al., 2019), whereas an SVM model when trained on a set of relevant features requires less data and is also well suited for binary classification (Cortes and Vapnik, 1995).

7 Conclusion and Future Work

This paper sought to determine whether binary comparative LCP was possible when attempting to differentiate the complexities of two different target words in the same sentence. Only our SVM and BERT + MLP models were found to be successful having achieved F1-scores of 0.86 and 0.81 respectively. This led to the conclusion that our SVM and BERT + MLP models benefited from more varied word-level feature representations of target word only input than in comparison to less varied contextual input. We also believe that more data is required to conduct further experimentation and achieve greater performances, especially with transformer-based models. The CompLex-BC dataset will be made freely available to the research community after the publication of this manuscript.

We are currently working on exploring different variables that may impact the modeling of binary comparative LCP. This includes evaluating the performance of models on target words in different contexts as well as exploring a "neutral" class with words with similar complexity scores.

Finally, we are interested in investigating the feasibility of binary comparative LCP on languages other than English. At this point, the 5-point likert-scale annotation introduced by CompLex 2.0 is only available in English, however, we expect multilingual versions of CompLex 2.0 to become available soon enabling us to work on languages other than English.

Acknowledgements

We would like to thank the anonymous BEA reviewers for their insightful feedback and suggestions.

References

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg, Sweden.

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):161.

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.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of AAAI-98 Workshop*.

Prafulla Choubey and Shubham Pateria. 2016. Garuda & Bhasha at SemEval-2016 Task 11: Complex Word Identification Using Aggregated Learning Models. In *Proceedings of SemEval-2016*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

BNC Consortium. 2007. British national corpus, XML edition. Oxford Text Archive.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(1):273–297.

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

Abhinandan Desai, Kai North, Marcos Zampieri, and Christopher Homan. 2021. LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction. In *Proceedings of SemEval-2021*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of NAACL*.

Hansi Hettiarachchi and Tharindu Ranasinghe. 2021. Transwic at semeval-2021 task 2: Transformer-based multilingual and cross-lingual word-in-context disambiguation. In *Proceedings of SemEval-2021*.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of ROCLING 2013*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Onur Kuru. 2016. AI-KU at SemEval-2016 Task 11: Word Embeddings and Substring Features for Complex Word Identification. In *Proceedings of SemEval-2016*.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of EMNLP*.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of SemEval-2016*.

Neil P. Morris, Mariya Ivancheva, Taryn Coop, Rada Mogliacci, and Bronwen Swinnerton. 2020. Negotiating growth of online education in higher education. *International Journal of Educational Technology in Higher Education*, 17(48):1–16.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of SemEval-2016*.

Gustavo Paetzold and Lucia Specia. 2017. Lexical simplification with neural ranking. In *Proceedings of EACL*.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 Task 1: Lexical Complexity Prediction with A Deep Ensemble Approach. In *Proceedings of SemEval-2021*.

Alice Pintard and Thomas François. 2020. Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of READI*.

Gang Rao, Maochang Li, Xiaolong Hou, Lianxin Jiang, Yang Mo, and Jianping Shen. 2021. RG PA at SemEval-2021 Task 1: A Contextual Attention-based Model with RoBERTa for Lexical Complexity Prediction. In *Proceedings of SemEval-2021*.

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 INTERACT*.

S.P Sanjay, Kumar M. Anand, and K.P Soman. 2016. AmritaCEN at SemEval-2016 Task 11: Complex Word Identification using Word Embedding. In *Proceedings of SemEval-2016*.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of READI*.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of SemEval-2021*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: The complex 2.0 dataset. *Language Resources and Evaluation*.

Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pre-trained Language Models. In *Proceedings of SemEval-2021*.

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 BEA*.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of NLPTEA*.

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 SemEval-2016*.