Manchester Metropolitan at SemEval-2021 Task 1: Convolutional Networks for Complex Word Identification

Robert Flynn
School of Computing, Mathematics
and Digital Technology
Manchester Metropolitan University
robert.flynn@stu.mmu.ac.uk

Matthew Shardlow
School of Computing, Mathematics
and Digital Technology
Manchester Metropolitan University
m.shardlow@mmu.ac.uk

Abstract

We present two convolutional neural networks for predicting the complexity of words and phrases in context on a continuous scale. Both models utilize word and character embeddings alongside lexical features as inputs. Our system displays reasonable results with a Pearson correlation of 0.7754 on the task as a whole. We highlight the limitations of this method in properly assessing the context of the target text, and explore the effectiveness of both systems across a range of genres. Both models were submitted as part of LCP 2021, which focuses on the identification of complex words and phrases as a context dependent, regression based task.

1 Introduction

Complex Word Identification (CWI) involves identifying words that the reader may find difficult to understand. A word's complexity can depend on many factors and differ according to context. Further, assessment of the complexity of named entities can require some degree of general knowledge, making CWI a challenging task (Shardlow, 2013). Accurately identifying complex words is important for many downstream simplification tasks, making literature more accessible for people with conditions such as dyslexia (Rello et al., 2013), and the assessment of a text's readability as a whole (Dubay, 2004).

Our methodology plans to extend on previous convolutional network based approaches to CWI (Aroyehun et al., 2018; Sheang, 2019). With the goal of producing a system that is able to better model the complexities of phrases and unfamiliar words, within the English language.

Previous shared tasks on CWI addressed the problem as a binary and probabilistic classification type task, although human judgements on word complexity are not binary and exist on a continuous scale. Lexical Complexity Prediction (LCP) 2021 tries to address this and uses an augmented version of CompLex (Shardlow et al., 2020), a dataset annotated with a 5-point Likert scale. CompLex also features context-specific annotation, with words receiving different annotations depending on their context. The dataset provides annotations from three different domains: Bible, Biomed and Europarl (Shardlow et al., 2021).

The code for this task is available on GitHub[1].

2 Related Work

Word frequency is a commonly used feature in CWI (Gooding and Kochmar, 2018; Kajiwara and Komachi, 2018); words that appear frequently in language are more likely to be recognised and understood by the reader (Carroll et al., 1998). For the purpose of identifying medical terminology that may be unfamiliar to the lay reader, Elhadad (2006) leveraged lexical frequencies while also exploring the potential of other features such as word familiarity ratings from the MRC Psycholinguistic Database (Coltheart, 1981).

More recently lexical and psycholinguistic features have been utilized by machine learning tools, resulting in improved accuracy on these tasks. Through the use of an enseble-based voting method the CAMB system (Gooding and Kochmar, 2018) achieved state-of-the-art results in the 2018 CWI shared task (Yimam et al., 2018), employing a total of 27 lexical, morphological and psycholinguistic features. The CAMB system however does not consider the target words context, opting for a "greedy" approach towards phrase classification, marking all phrases as complex.

Aroyehun et al. (2018) explored the use of convolutional neural networks (CNN) for CWI using only

[1] https://github.com/robflynnyh/CNN-LCP-Shared-Task-2021

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 603–608
Bangkok, Thailand (online), August 5–6, 2021.

the word embeddings of the target words and the averaged embeddings of the left and right contexts as inputs. They contrasted the results against a feature engineering approach using decision tree learning finding that both methods achieved competitive results. However, their decision tree method was marginally more accurate than their CNN for most of the datasets. Integrating lexical features alongside word embeddings can lead to further improvements in accuracy making this a more competitive approach, and outperforming many previous deep learning methods for CWI (Sheang, 2019).

By framing CWI as a sequence labelling task, Bi-directional long short-term memory (BiLSTM) networks have produced state-of-the-art results on the CWIG3G2 dataset (Yimam et al., 2017; Gooding and Kochmar, 2019). BiLSTM networks are able to capture long-term word and character level dependencies allowing these models to consider a large amount of contextual information. Modelling the complexity of phrases has proven to be a more challenging and complex task compared to individual words (Gooding and Kochmar, 2019).

3 Implementation

3.1 Features

Below a description of the features used by both models is given:

Frequency: Word frequencies are taken from the SUBTLEX-UK word frequency database (van Heuven et al., 2014). Logarithmic Zipf frequency values were chosen based on previous results from this metric (Zampieri et al., 2016) and the Zipfian distribution that is displayed in language (Zipf, 1949).

Age of Acquisition: Age of Acquisition (AoA) values, estimating the age at which a word is typically acquired. (Kuperman et al., 2012; Brysbaert, 2012).

Word-level Features: Target word length and number of syllables are used as features (Brysbaert, 2012).

Corpus Type: As the dataset includes extracts from three different sources of potentially varying complexity, the corpus type was included and represented as a one-hot embedding.

Pre-trained Embeddings: 50d GloVe (Pennington et al., 2014) word embeddings, and 50d chars2vec[2] embeddings representing a word's character sequence are used. 50d GloVe embeddings were chosen as embeddings with more dimensions showed worse performance on the training data. Which suggests that the 50d embeddings capture sufficient information needed for this task. Character embeddings allow inferences to be made between words with similar morphologies.

[2] https://github.com/IntuitionEngineeringTeam/chars2vec

3.2 Preprocessing

Firstly min-max normalization is applied to the features taken from datasets, and word length is divided by 10. Non-alphanumeric characters are removed from the sentences before any features are extracted.

Both models take as inputs the features for the target word, and the averaged features for the left and right contexts of the target text. If the target word or words are positioned at the beginning or end of the sentence a zero vector of size 107 is used for the left or right context. For out-of-vocabulary words a zero vector is used for the word embedding and other features are imputed using mean values from their respective datasets. Finally the vectors for the target text and its context are stacked to produce a 3x107 matrix (left context — token — right context) for single words or a 4x107 matrix for MWEs (left context — token 1 — token 2 — right context).

3.3 Models

This section provides a description of the architecture and hyperparameters used for both models. The models were produced using the Keras library version 2.4.3. Each of the models were trained with a batch size of 50, early stopping of 1000 and model checkpointing based on the validation loss.

3.3.1 Single Word Model

For single words a 1D convolutional network followed by three fully connected layers is used. The model takes three inputs, an average of the features for left and right contexts is used for the first and third inputs respectively, and the features of the target word is used as the second input. The convolutional layer pads the inputs and uses a kernel size of 3 with 150 output filters and ReLu as the activation function. Global Max Pooling and a flatten layer followed by batch normalization is then applied to the output of this layer. Three dense layers with sizes of 150 (ReLu), 50 (ReLu) and 1 (Linear) are then used with a Dropout of 0.5 applied before each dense layer. Mean Squared Error

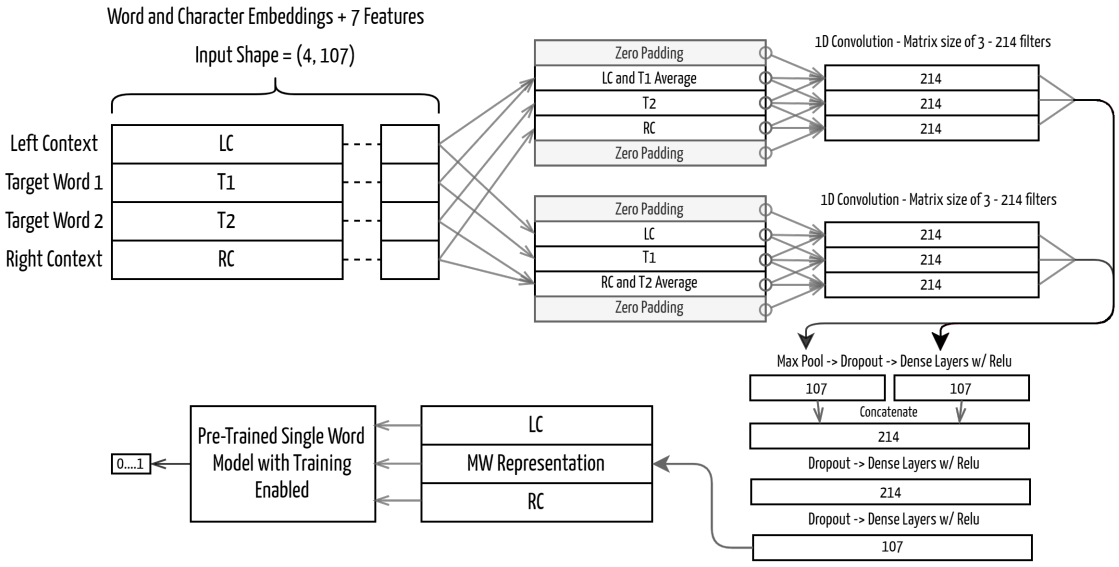

Figure 1: Depiction of multi-word model architecture

(MSE) is used as the loss function and Stochastic Gradient Descent as the optimizer, with a learning rate of 0.01 and momentum of 0.6 with Nesterov accelerated gradient enabled.

3.3.2 Multi-Word Model

For multi-words a second model is used to assess the complexity of two word phrases. This model acts as an adapter with the output being fed into a pre-trained single word model, allowing the model to take advantage of the data for single words and MWEs. Figure 1 gives an overview of the model architecture.

Features for the averaged left context, target word one, target word two and the averaged right context are used as input for the model. A convolutional layer with a similar architecture to task one is used for each of the target words. For the two convolutional layers the other target word is averaged with either the left or right context depending on its positioning, weighting the other target word higher than the rest of the context.

Each convolutional layer uses a filter size of 214 but is otherwise the same as in task one. Global Max Pooling followed by Dropouts of 0.3 and dense layers with 107 neurons and ReLu activation are applied to the outputs of the convolutions which are then concatenated along the last axis. Two dense layers with ReLu activation and sizes of 214 and 107 are then applied with a Dropout of 0.5 before each layer. This final output of size 107 is then concatenated along the first axis with the original left and right contexts to form the input for a pre-trained single word model with training enabled. This model uses the Adam optimizer with default parameters and MSE as the loss function.

4 Results

Task	Pearson	MSE	R2
Task 1	0.7389	0.0074	0.5398
Task 2	0.7754	0.0079	0.5989

Table 1: Results for both tasks

This section will discuss and evaluate the performance of both models. Participants were ranked according to the Pearson correlation coefficient of their submissions. Table 1 presents the results for each of the tasks with task 1 evaluating individual words and task 2 covering both single and two word Multi-Word Expressions (MWEs).

4.1 Single Word Model Results

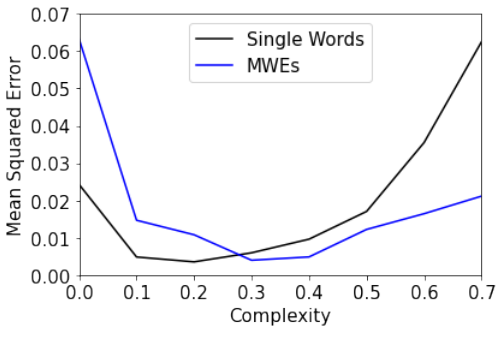

Figure 2: MSE across different complexities

As shown in Figure 2 the single word model struggles to accurately predict values for words of a high complexity, and also displays difficulties for words of a complexity of less than 0.1. The training and evaluation data features less examples of very simple or complex words. The complexity of these extremities is often highly dependant on the context, making them more challenging to assess.

Corpus	Pearson	MSE	R2
All	0.7389	0.0074	0.5398
Bible	0.7085	0.0085	0.4948
Biomed	0.7828	0.0087	0.6050
Europarl	0.6807	0.0055	0.4562
JUST-BLUE	0.7886	0.0062	0.6172

Table 2: Results for individual words

Table 2 presents the results for this task on each of the domains and the task as a whole. The prediction accuracy varies significantly across the different sources. Results from the best performing team are given for comparison (Shardlow et al., 2021).

As the model only uses an average of the features present in the left and right context of the target word, it is unable to differentiate between tokens that are influential to the target words complexity and ones that are not. Because of this equal weighting of words in the context, the models accuracy can be negatively affected by an abundance or lack of stop words in the sentence. Very complicated or simple words in sentence that are not related to the target word, and don't share a similar complexity can also cause the model to over- or under-predict the target word's complexity. The mechanism by which the model assesses the context may partly explain the variance in accuracy on each domain.

Interestingly, our sub-analysis showed that the model shows a better correlation for those tokens without a word embedding, yielding a Pearson correlation of 0.7804 and a MSE of 0.0071. Generally these out-of-vocabulary words are more complex so the model is using the lack of a word embedding as a feature when making predictions. Although this shows a better correlation overall it could lead to false positives in specific cases where the out-of-vocabulary word is of a low complexity.

4.2 Multi-Word Model Results

As shown in Figure 2 the multi-word model is much less accurate for very simple MWEs of a complexity less than 0.1. However, for more complex words the predictions remain fairly accurate. This model is able to asses the way in which the words in a phrase interact with each other and to some degree the rest of the sentence. This additional contextual information may increase the model's capacity to evaluate more complex words. Only 1.65 percent of phrases in the training data were of a complexity of less than or equal to 0.1 which could explain the inaccuracy in this range.

Corpus	Pearson	MSE	R2
All	0.7611	0.0102	0.5770
Bible	0.7173	0.0113	0.5106
Biomed	0.7980	0.0141	0.6317
Europarl	0.5799	0.0060	0.3089
DeepBlueAI	0.8612	0.0063	0.7389

Table 3: Results for MWEs

MWE Type	Pearson	MSE	R2
A-N (115)	0.7654	0.0115	0.5801
N-N (56)	0.7414	0.0091	0.5293

Table 4: Results for the different MWE formations. A-N: Adjective-Noun. N-N: Noun-Noun.

Table 3 presents the results across each of the different domains present in the dataset. The model used for MWEs makes use of a fine-tuned instance of the single-word model; consequentially incorrect associations from the single-word model may have been carried over to this model. The results show a similar variance across domains to task 1, although it struggles more significantly on the Europarl subcorpus. Compared to the other domains, Europarl's complexity values show a much smaller standard deviation than the other sub-corpora (0.093 compared to 0.196 and 0.152, on biomed and bible). The variation of complexities may play a role in the models effectiveness at making accurate predictions across the domains.

Table 4 presents the results across different MWE formations. The number of occurrences of each part-of-speech formation is denoted in brackets, MWE types with less than 10 occurrences were omitted from the table. The model performs marginally better on Adjective-Noun MWE formations.

5 Discussion

In this paper, we presented two convolutional neural networks used as an approach to single-word and multi-word complex word identification. Both models achieved reasonable results, achieving scores in a comparable range to the majority of other submissions.

Multi-Word CWI is a more challenging task compared to the assessment of single words; the multi-word model was able to utilize the datasets of both tasks, and its predictions show a Pearson's correlation score of 0.7611. Our system is only able to process two-word MWEs, which for this task is not an issue. However, in other use cases the ability to assess longer MWEs would be useful. Given a dataset with annotations for longer MWEs the model could potentially be adapted to work with three or four word sequences.

Both models are able to assess the context of the target text when making predictions; although, as the left and right contexts are given as an average, all words are weighted equally regardless of their relevance to the target text. Because of this equal weighting of words, the models are able to adjust their predictions based on the general complexity of the sentence but are unable to fully capture the relevant context. Adding a mechanism that could weight each word in the context based on certain features may offer some improvements in this area. Attention based models such as BERT (Devlin et al., 2019) are able to attend to each token in a sequence to produce embeddings that capture large amounts of contextual information. Fine-tuning such a model on CWI tasks could produce embeddings that contain more useful and relevant information.

References

Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, and Alexander Gelbukh. 2018. Complex word identification: Convolutional neural network vs. feature engineering. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 322–327, New Orleans, Louisiana. Association for Computational Linguistics.

Marc Brysbaert. 2012. crr " age-of-acquisition (aoa) norms for over 50 thousand english words.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dubay. 2004. The principles of readability. *CA*, 92627949:631–3309.

Noemie Elhadad. 2006. Comprehending technical texts: predicting and defining unfamiliar terms. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2006:239–243. 17238339[pmid].

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.

Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Walter van Heuven, Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: a new and improved word frequency database for british english. *Quarterly journal of experimental psychology (2006)*, 67.

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.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA. Association for Computing Machinery.

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 prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

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

Kim Cheng Sheang. 2019. Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 83–89, Varna, Bulgaria. INCOMA Ltd.

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.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. MacSaar at SemEval-2016 task 11: Zipfian and character features for ComplexWord identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1001–1005, San Diego, California. Association for Computational Linguistics.

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*.