SINAI at SemEval-2021 Task 1: Complex word identification using Word-level features

Jenny Ortiz-Zambrano
Universidad de Guayaquil
Guayaquil, Ecuador
jenny.ortizz@ug.edu.ec

Arturo Montejo-Ráez
CEATIC - Universidad de Jaén
Jaén, España
amontejo@ujaen.es

Abstract

This article describes a system to predict the complexity of words for the Lexical Complexity Prediction (LCP) shared task hosted at SemEval 2021 (Task 1) with a new annotated English dataset with a Likert scale. Located in the Lexical Semantics track, the task consisted of predicting the complexity value of the words in context. A machine learning approach was carried out based on the frequency of the words and several characteristics added at word level. Over these features, a supervised random forest regression algorithm was trained. Several runs were performed with different values to observe the performance of the algorithm. For the evaluation, our best results reported a M.A.E of 0.07347, M.S.E. of 0.00938, and R.M.S.E. of 0.096871. Our experiments showed that, with a greater number of characteristics, the precision of the classification increases.

1 Introduction

The identification of complex words (CWI) is the task of detecting in the content of documents the words that are difficult or complex to understand by the people of a certain group (Rico-Sulayes, 2020). The CWI and the substitution of words identified as complex may significantly improve readability and understandability of a given text (Zotova et al., 2020).

CWI has become an area of great interest in recent years for the computational linguistics community in making proposals that allow researchers to develop computational semantic analysis systems, as demonstrated by the shared tasks of CWI in SemEval 2016 (Paetzold and Specia, 2016), y NAACL-HTL 2018 (Yimam et al., 2018), and the CWI task of the ALexS 2020 competition, hosted at IberLEF 2020 (Ortiz-Zambranoa and Montejo-Ráezb, 2020).

This article introduces a system that has participated in the Lexical Complexity Prediction (LCP) shared task hosted at SemEval 2021 (Task 1) (Shardlow et al., 2021a). The task releases a new annotated English dataset with a Likert scale. Located in the Lexical Semantics track, the task consisted of predicting the complexity value of the words in context.

We have explored different features for representing words and multi-words and their context. Some preprocessing steps have been evaluated along with the effect of feature selection.

2 Related Work

(DuBay, 2004) defines readability as allowing one text to be easier to read than another. For many people, the understanding of a text can be affected by the presence of lexically and semantically complex words and phrases, for example for children (Petersen and Ostendorf, 2009), non-native speakers (Petersen and Ostendorf, 2009), and people with various cognitive or reading disabilities (Saggion et al., 2015).

Predicting which words a given target population has difficulty to understand is a critical step for many NLP applications, such as in text simplification, which has traditionally focused its attention on second language learners, native speakers with low levels of literacy, and people with language disabilities reading (Saggion et al., 2015). This task is also known as complex word identification (CWI). The prediction of the lexical complexity carried out with precision can allow to adapt texts according to the needs of the readers (Shardlow et al., 2020). Actually, in an early study in the 1920s, a very simple way to predict the level of difficulty of a text was discovered by educators, who used vocabulary difficulty and sentence length as main indicators (DuBay, 2004).

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 126–129
Bangkok, Thailand (online), August 5–6, 2021.

corpus	bible	europarl	biomedic	total
single	2574	2576	2512	7662
multi	505 1	498	514	1517

Table 1: Total number of sentences in each training corpus.

3 Dataset

The training data set provided to the participants consisted of an augmented version of CompLex (Shardlow et al., 2021b). It uses data from three different sources: the Bible, Europarl, and biomedic texts (see Table 1). It is a set of multidomain English data made up of sentences, the targeted token, and its respective level of complexity as described in (Shardlow et al., 2020).

4 The system

This section describes the details of the system applied to the task, as our approach to complex word identification. A machine learning approach was followed based on the frequency of the words and further characteristics added at word level. Over these features, a supervised random forest regression algorithm was trained. In this section, first, the features considered in the supervised learning approach are introduced. Then, the method to determine whether a candidate word is complex or not is detailed.

4.1 Features

We computed a total of 15 features, taking into consideration the linguistic measures of the work carried out by (Mc Laughlin, 1969) and the experiments of the shared tasks of the CWI BEA 2018 respectively by (Paetzold and Specia, 2016; Gooding and Kochmar, 2018). These are the features obtained on the target word (token).

- *Absolute frequency (abs-frequency)*: the absolute frequency. This frequency is computed based on the unannotated corpora compiled by José Cañete[1] from different sources. It contains about 3 billion words.
- *Relative frequency (rel-frequency)*: the relative frequency of the target word.
- *Word length (length)*: the number of characters of the token.
- *Number of syllables (number-syllables)*: the number of syllables.
- *Target word position (token-position)*: the position of the target word in the sentence.
- *Number of words in the sentence (n-words-sentences)*: number of words in sentence.
- *Part Of Speech (POS)*: the Part Of Speech category.
- *Relative frequency of the previous the token (freq-rel-word-before)*: the relative frequency of the word before the token.
- *Relative frequency of the word after the token (freq-rel-word-after)*: the relative frequency of the word after the token.
- *Length of previous word (len-word-before)*: the number of characters in the word before the token.
- *Length of the after word (len-word-after)*: the number of characters in the word after the token.
- *Measure of Textual Lexical Diversity (MTLD-diversity)*: the lexical diversity of the target word in the sentence using the metric proposed by (McCarthy and Jarvis, 2010)[2].

Additionally, the following WordNet (Fellbaum, 2010) features were also considered for each target word:

- *Number of synonyms (number-synonyms)*.
- *Number of hyponyms (number-hyponyms)*.
- *Number of hyperonyms (number-hypernyms)*.

In the case of multiple words, the following characteristics were applied: absolute frequency, relative frequency, token length, number of syllables, total number of words in the sentence, MTDL diversity.

[1] Available at https://github.com/josecannete/spanish-corpora

[2] Computed using this Python library: https://pypi.org/project/lexical-diversity/

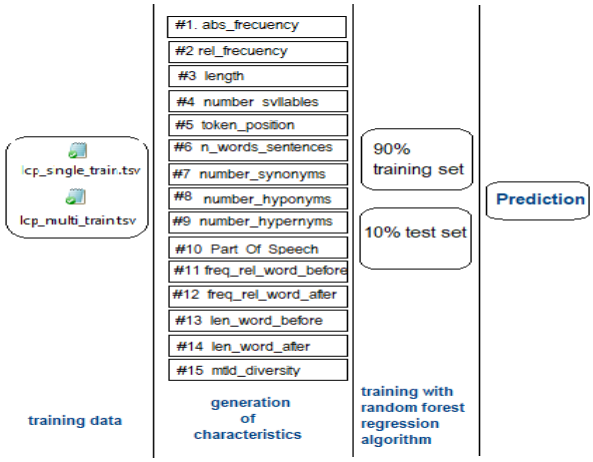

Figure 1: Training process applying the Random Forest Regression algorithm. A different model is trained for each training subset of data.

4.2 Method

The numeric input variables were scaled to a standard range, as many machine learning algorithms have been found to perform better when the data set is normalized. A polynomial transformation on the features characteristics was then applied with a degree value of 2, so new features were created.

A forest of trees was built with the training set (X, y), where we assigned to the independent variable (X) an array that contains all the word-level characteristics that were obtained from the token, the same ones that were described in the section 4.1; and the value of the dependent variable (y) corresponds to the level of complexity's word.

To build the Random Forest Regression Model, we split the dataset into the training set and test set, that is, 10% of the data set was used as test set, and the remaining 90% was used as the training set. Figure 1 shows the training process applying the random forest regression algorithm.

5 Experimental Results

5.1 Results on Trial and Simulated Data

To calculate the prediction value of the word complexity on the data of the evaluation corpus, the (X, y), where we assigned to the independent variable (X) which we called XTest, was built, was an array that contained all the word-level characteristics that were obtained from the token. Finally, we train the algorithm with the evaluation data and predict the results of the test set with the model trained on the testing set values using the regressor predict function.

#Trees	K	MAE	MSE	RMSE
150	7	0.07347	0.00938	0.09687
130	7	0.07354	0.00940	0.09700
150	8	0.07356	0.00942	0.09710

Table 2: Results obtained with Random Forest with selecting K-best features on single words subset.

#	Team Name	MAE	MSE	R^2
1	JUST_Blue	0.0609	0.0062	0.6172
2	DeepBlueAI	0.0610	0.0061	0.6210
3	Alejandro M.	0.0619	0.0064	0.6062
50	SINAI	0.0875	0.0131	0.1930

Table 3: Final results of the Lexical Complexity Prediction task on the single words dataset

Several runs were made with different values to observe the performance of the algorithm and fine-tune the hyperparameters of the model.

Our best configuration was with 150 nodes and 7 features, selected by their F ANOVA between label / feature. The selected characteristics were: abs_frecuency, rel_frecuency, lenght, number_syllables, token_position, number_synonyms, Part_of_speech. Finally, the prediction value of the words for the test data set was obtained, obtaining the best result: MAE of 0.07347, MSE of 0.00938, and RMSE of 0.096871 (see Table 2).

5.2 Results on test Data

In this section we present the results obtained from our system, and we carry out a discussion regarding the results presented by the organizers of the workshop.

The final results were sent to the SemEval 2021 organizers after the execution of our system. The final published results are those shown in Table 3, where the winners of the first three positions are presented. The results that we obtained in the contest for the case of the evaluation corpus of simple words were, MAE of 0.0875, MSE of 0.0131 and R-squared of 0.1930. Taking into account the number of competitors (quite large) and the result obtained by the first place winner (MAE of 0.0609), we see that there is a small difference, which allows us to be confident with our simple approach.

6 Conclusion

In this article, the results of our participation in Task 1: Lexical Complexity Prediction in the Lexical semantics track hosted at the SemEval 2021 international workshop have been presented. Both the training corpus and the evaluation corpus were provided by the sponsoring organization of this competition. We applied machine learning and built the model using the random forest regression algorithm, relying on well-known word based and contextual features.

As future work, we plan to perform error analysis on the predictions, to identify the weaknesses of the proposed approach based on a characterization of the instances where the system performs poorly. Also, a better analysis of multi-word scenario is foreseen.

Acknowledgments

We appreciate , Luis Alexander Suárez Colimba, graduates of the Computer Systems Engineering degree from the University of Guayaquil, for their valuable contribution to the development of our work.

This study is partially funded by the Spanish Government under the LIVING-LANG project (RTI2018-094653-B-C21).

References

William H DuBay. 2004. The Principles of Readability. *Online Submission*.

Christiane Fellbaum. 2010. WordNet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

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.

G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Jenny A Ortiz-Zambranoa and Arturo Montejo-Ráezb. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. pages 560–569.

Sarah Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech Language*, 23:89–106.

A Rico-Sulayes. 2020. General Lexicon-Based Complex Word Identification Extended with Stem N-grams and Morphological Engines. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput.*, 6(4).

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. 2021a. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021)*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2021b. Predicting Lexical Complexity in English Texts. *arXiv preprint arXiv:2102.08773*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. *arXiv preprint arXiv:1804.09132*.

Elena Zotova, Montse Cuadros, Naiara Perez, and Aitor García-Pablos. 2020. Vicomtech at ALexS 2020: Unsupervised Complex Word Identification Based on Domain Frequency. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*.