CompNA at SemEval-2021 Task 1: Prediction of lexical complexity analyzing heterogeneous features

Giuseppe Vettigli
Centrica plc
giuseppe.vettigli@centrica.com

Antonio Sorgente
Institute of Applied Sciences
and Intelligent Systems
National Research Council
antonio.sorgente@isasi.cnr.it

Abstract

This paper describes the CompNa model that has been submitted to the Lexical Complexity Prediction (LCP) shared task hosted at SemEval 2021 (Task 1). The solution is based on combining features of different nature through an ensembling method based on Decision Trees and trained using Gradient Boosting. We discuss the results of the model and highlight the features with more predictive capabilities.

1 Introduction

Complex Word Identification (CWI) is a task focused on the detection of complex (not easy to understand) words or expressions. One of the challenges of natural language-based systems is to provide informative content that is tailored to the needs of users in terms of content and level of understanding. For this aim, predicting Lexical Complexity plays a crucial role in simplifying the text so that it can be more easily understood by people with low literacy levels (for example children) or non native speakers (Shardlow, 2013). The interest of the Computational Linguistics community on this topic has grown in recent years. Indeed, SemEval 2016 presented a challenge specifically on CWI (Paetzold and Specia, 2016a; Zampieri et al., 2017). The task proposed to build models capable of predicting whether a word was easy or not for non-native English speakers. In this case, the proposed dataset was annotated with a binary label.

In 2018 the CWI task was re-proposed at the workshop for Building Educational Applications (BEA) (Yimam et al., 2018). The data was again annotated with binary labels but samples from more languages were considered, specifically English, German and Spanish. In addition to the word classification task, a sub-task where the participants had to predict the probability of a word being complex was added.

The best performing models for both tasks were based on ensembling techniques using features that were carefully selected (Paetzold and Specia, 2016b; Gooding and Kochmar, 2018).

In this context, the Lexical Complexity Prediction (LCP) shared task hosted at SemEval 2021 proposes a challenge where the data is annotated according to the degree of complexity (Shardlow et al., 2021). This type of annotation allows regression models that predict a complexity index rather than just a binary label. This task is split into two sub-tasks, the first one is about the prediction of the complexity of single words and the second is about the prediction of the complexity of multi word expressions.

In this work we present our submission to both the sub-tasks with a model that aggregates a large set of heterogeneous features that can capture a wide variety of linguistics aspects (morphological, semantic, distributional and lexicon-based) in a regression model based on Gradient Boosting. We also present an in-depth analysis of which features are more important for the model.

In Section 2 we present a description of the task and the data available. In Section 3 we introduce the model. In Section 4 we show an analysis of the most relevant features for the model. In Section 5 we present the results. Finally, in Section 6 we draw some conclusions.

2 Task and data

The data released for sub-task 1 is made of 9000 sentences where the complexity of a single word was annotated considering its context. The complexity has been annotated using a 5-point Likert scale (from 1 to 5 corresponding to Very Easy, Easy, Neutral, Difficult and Very Difficult) with the values scaled in the range 0-1. Here is one example of a sentence annotated in the dataset:

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 560–564
Bangkok, Thailand (online), August 5–6, 2021.

> "The structural Gh gene itself and Stat5b are excellent candidates."

In this sentence, the word "candidates" is annotated with a complexity of 0.11 (the average in the dataset is 0.30). The goal of sub-task 1 is to predict the complexity of a single word given the sentence and the word to evaluate. The data for sub-task 2 is made of 1800 sentences where expressions of two words are annotated as already described. Here is one sample extracted from the dataset:

> "I invite the President to ask all Members who are in the pension fund to say so orally, in plenary, immediately, because they have a direct interest in what is to be discussed."

In this sentence, the expression "direct interest" is annotated with a complexity of 0.40 (the average in the dataset is 0.42). The goal of sub-task 2 is to predict the complexity of an expression like the one above given the sentence and the expression to evaluate.

For both sub-tasks, the data was selected from three different corpora:

- The World English Bible, translation from Christodouloupoulos and Steedman (2015).
- A selected portion of the European Parliament proceedings in English.
- Selected articles from the biomedical domain.

Having data from sources so diverse is a unique aspect of this task. This dataset was introduced in (Shardlow et al., 2020).

3 CompNA model

The main idea behind our model is to aggregate many diverse features that can capture a wide variety of linguistics aspects in a regression model that offers the ability to interpret which of them are more influential.

3.1 Features

In this section we list the features used for the two sub-tasks.

3.1.1 Features for sub-task 1

The following sets of features were used to capture morphological aspects of the text:

- Part of speech tag and Syntactic dependency of the word to evaluate and surrounding words in. a window of three words. The library Spacy with the model en_core_web_sm (Honnibal and Montani, 2017) has been used to extract these features.
- Syllables in the word. Number of syllables in the word. Minimum, maximum and average number of syllables in the sentence.
- Length of the word. Minimum, maximum, and average of the lengths of all the words in the sentence.
- Desinence of the word, first and last letter of the word.
- Number of unique characters.
- A Boolean value that is 1 only if all the characters in the word are uppercase.

To take into account distributional characteristic of the word to evaluate we used:

- GloVe embedding of the word, we used the version pre-trained on Wikipedia 2014 and Gigaword 5 with size 50 (Pennington et al., 2014).
- Frequency of the word and of the part of the sentence of three words that include the word in the wordfreqs dataset (Speer et al., 2018).

Semantic aspects were encoded considering the number of synsets and hyponyms of the word in WordNet.

We also considered a set of binary features that report if the word is in one of the following lexicons: Obscure words (Chrisomalis, 1996), Medical words, Simple English words (Ogden and Halász, 1935).

3.1.2 Features for sub-task 2

For sub-task 2 we have used the same features considered for sub-task 1 computing them for each word in the expression to evaluate. We have also restricted the part of speech tags to the target words and added the frequency of the entire expression.

3.2 Regressor

Our final regressor is composed of 120 Decision Trees with a maximum depth of 5 layers. The model was trained using Gradient Boosting (Friedman, 2001) with a learning rate of 0.047 and taking a sub-sample of 75% of the original data at each boosting iteration. The final prediction is computed averaging the output of each tree. The library XGBoost (Chen and Guestrin, 2016) was used for our experiments. The parameters were selected via a grid search performed using the library Scikit-Learn (Pedregosa et al., 2011).

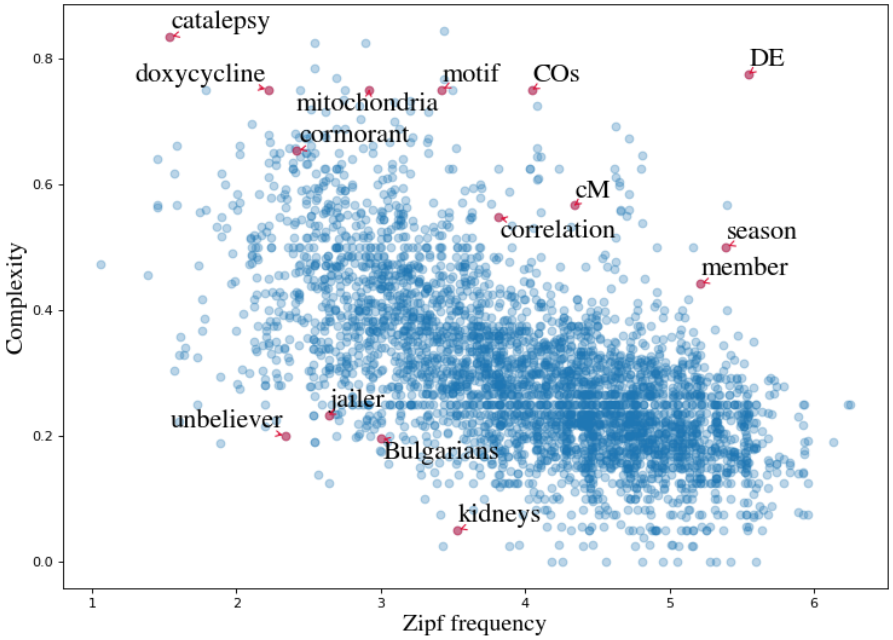

Figure 1: Frequency of the word in Zipf scale versus the complexity. The points in red represent words for which the model has an error higher than a quarter of a point. The chart was made splitting the train data for sub-task 1 in half. The first half has been used to train the model and the second to evaluate the error.

Features group	Importance	N. features
Syllables	19.1%	2260
GloVe	17.8%	50
Desinences	10.2%	898
All frequencies	8.7%	6
Starting letters	5.8%	52

Table 1: Importance of the top 5 groups of features for the model (on single words).

Feature	Importance	Type
All uppercase letters	9.47%	binary
Word frequency	7.93%	float
In medical lexicon	3.03%	binary
POS PROPN	1.85%	binary
Word length	1.47%	integer
Syllable “ca”	1.19%	binary
Ends with “s”	1.10%	binary
Desinence “ess”	0.84%	binary
Ends with “e”	0.75%	binary
Desinence “ing”	0.71%	binary
GloVe 32th position	0.68%	float
Syllable “ver”	0.67%	binary
Syllable “ro”	0.64%	binary
Desinence “ium”	0.63%	binary
GloVe 29th position	0.62%	float

Table 2: Importance of the top 15 single features for the model (on single words).

4 Features importance

In order to bring insights into how the model works, we study the contribution of the features. Tables 1 and 2 report the importance of the features that are influential for the prediction, respectively, per group and single feature. The feature importance reflects the number of times a feature is selected for a split for one of the decision trees in the model, weighted by the improvement as a result of the split, and averaged over all the trees in the model (Elith et al., 2008).

Looking at the importance of the features grouped by type we find morphological and distributional ones at the top. While the morphological group (syllables, desinences and starting letters) add up to almost 3000 features, the distributional group, made of GloVe and word frequencies, only presents 56 features.

Going into details and inspecting the importance

Sub-task	Team	Pearson	Spearman	MAE	MSE	R2
1	JUST Blue	0.7886	0.7369	0.0609	0.0061	0.6172
1	CompNa	0.7552	0.7153	0.0641	0.0070	0.5701
1	*baseline*	0.6920	0.6533	0.0737	0.0091	0.4387
2	DeepBlueAI	0.8612	0.8526	0.0616	0.0063	0.7389
2	CompNa	0.7931	0.7800	0.0783	0.0093	0.6160
2	*baseline*	0.7503	0.7435	0.0848	0.0111	0.5386

Table 3: Results of the CompNA model compared a baseline and the best performing model in the competition. The baseline is given by the values on the 20th percentile of the leaderboard.

of single features we note that two of them stand out covering more than 15% of importance. The first is the binary variable representing if the word to evaluate is made of only capital letters and the second is the frequency of the word. In Figure 1 we see the frequency of the word in Zipf scale versus the complexity annotated. The frequency of a word in Zipf scale is the base-10 logarithm of the number of times it appears per billion words (Van Heuven et al., 2014). The Figure shows in red the words for which the model achieves an error greater than 0.25. It is easy to see that the relation between the two features in the chart is well covered by the model as significant errors happen mostly in samples that can be considered outliers. Specifically, we see that many acronyms tend to be outliers in this space. It is interesting to note that acronyms highlighted in the chart are annotated with a score much higher than the average.

Looking again at the table of individual features importance, we find the desinence "ess" which is associated with simple words (such as "darkness", "kindness" and "business") and the desinence "ium" which is associated with words that have a complexity more than the average (such as "epithelium", "cadmium", "medium").

Considering the most important syllables, we notice the syllable "ca" which is associated with words from a wide range of complexities (the most complex "cause", and the least complex "catalepsy"). And the syllable "ver" which is associated with low complexity words (the most complex being "reversal", and the least complex being "river").

This analysis further confirms the hypothesis outlined in (Zampieri et al., 2016) that distributional and morphological aspects of the words have a tight bond with the complexity.

The analysis also highlights one of the weak points of the model, it overlooks features related to the context of the word.

5 Results

In Table 3 we summarize the results comparing our model with the best performing model in the competition and a baseline given by the values on the 20th percentile of the leaderboard. The proposed model largely outperforms the baseline in all the measures considered. Notice that the baseline considered here is more challenging than the one proposed in the description of the data (Shardlow et al., 2020). Regarding sub-task 1, the results of the proposed model are comparable to the ones of the winner of the competition in all the measures apart from the R^2 where the two models have a difference of almost 18%. While for sub-task 2 the model beats the baseline but it is far from the winning model. For sub-task 1 the model ranked 26th achieving above average performances and for sub-task 2 it ranked 24th achieving average performances.

6 Conclusions

In this paper, we presented a solution to predict the complexity of single and multi word expressions combining a large number of features from a diverse nature.

The model achieves average results and, more importantly, offers the ability to quantify the relevance of the features for the prediction. Thanks to this ability we have shown that the frequency of occurrence and the morphology of the words are key predictors of complexity for the data considered.

From our analysis, it is clear that our model overlooks features related to the context of the word and we would like to improve it under this point of view in our future efforts.

References

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Stephen Chrisomalis. 1996. The phrontistery: Obscure words and vocabulary resources. Available: http://phrontistery.info/.

Jane Elith, John R Leathwick, and Trevor Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

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.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Charles Kay Ogden and Gyula Halász. 1935. *Basic English*. Kegan Paul Trench Trubner. Available: http://www.basic-english.org/down/download.html.

Gustavo Paetzold and Lucia Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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.

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)*.

Matthew Shardlow, Marcos Zampieri, and Michael Cooper. 2020. Complex—a new corpus for lexical complexity predicition from likertscale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with READing DIfficulties (READI)*, pages 57–62.

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

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

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*.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*.

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.