Complexity as a regression task

Trung Hieu Ngo, Nicolas Béchet, Delphine Battistelli

Workshop "La complexité dans les sciences du langage", 12 et 13 décembre 2024, Maison de la Recherche, Paris

Complexity as a regression task

Trung Hieu Ngo (LS2N, CNRS-Nantes University), Nicolas Béchet (IRISA, CNRS-South Brittany University), Delphine Battistelli (MoDyCo, CNRS-Paris Nanterre University)

Keywords: complexity, regression, descriptors, Natural Language Processing

Our work aims to observe which are the most discriminating linguistic descriptors in order to produce an efficient regression model for text complexity in French. Text complexity is currently not an area with extensive studies in Natural Language Processing (NLP). Among the studies, there is research on age recommendation and text readability using classification (Mesgar and Strube, 2018; Balyan et al., 2020) and using regression (Bayot and Gonçalves, 2017; Chen et al., 2019). The regression approach shows promise over classification approach due to the more fine-grained nature of regression over classification.

Our work takes place in the ANR project TextToKids[1] which is a multidisciplinary project, combining experts from linguistics, psycholinguistics, and natural language processing (NLP). This project tackles the problem of childrens' - from 7 y. old et 12 y. old - access informational content of genre diversified texts (journalistic, fictional, encyclopedic). Thus, it is directly concerned with the question of how to evaluate complexity of texts for this type of population. For solving this question, the project focuses on how to describe complexity into elementary descriptors objectively. The project has made major achievements in creating in particular two automatic tools, one for extracting linguistic descriptors (Battistelli et al., 2022) and one for predicting recommended minimal age ranges for texts' readers (Rahman et al., 2020, 2023) starting from a corpus annotated in age ranges as proposed by publishers. The goal of this communication is to present the application of the combination of these two automatic tools on a dataset of pairs of texts: experts' manually simplified texts together with their original versions. This dataset is named Alector corpus (Gala et al., 2020) and we aim to prove that the combination of our two automatic tools is able (1) to detect in a pair of texts which one is the simplified one ; (2) to identify which descriptors are the most impactful descriptors for representing the difference in complexity between the original texts and their simplified versions.

As we said, Alector is a parallel corpus of original and simplified French texts. It is drawn from 79 French literary and scientific texts commonly used in schools for children from 7 to 9 years of age. The corpus is organized into age grade levels of CE1 level (7 y. old), CE2 level (8 y. old), and CM1 level (9 y. old), along with samples from International Reading Tests to enrich the corpus. The simplifications were manually done at the lexical, morpho-syntactic, and discourse level. To carry out our study, the linguistic features are first extracted from the original and the simplified versions using the extraction tool described in (Battistelli et al., 2022)[2]. This tool allows the extraction of a total of 20 descriptors from 6 groups, namely Phonetic, Morphology, Morphosyntax, Lexical, Syntax, and Semantic. Using interpretability tools such as SHAP (Lundberg and Lee, 2017), we can find which descriptors are more important for understanding text complexity. The extracted features are then used as input representation for our trained regression models such as XGBoost (Chen and Guestrin, 2016) to predict the recommended age for a given pair of original and simplified texts, and the interpretability module is used to show the most impactful descriptors that can highlight the difference in complexity between the two versions.

The results from our work show that: (1) based on the linguistic descriptors, the original version is indeed more complex than the simplified versions. Using the differences between predicted age ranges as a measure, we see that the original texts have on average +1.295 years higher recommended age than the simplified texts; (2) using the interpretability module, we extracted "niveau_lexical", "phonetique", "pronoms", "parties_du_discours", "dependances_syntaxiques", and "flexions_verbales" as the most impactful descriptors to describe text complexity. These descriptors correspond to "lexique", "phonetique", "morphosyntax", and "syntaxe" groups of descriptors, which are in line with the experts' manual simplification process reported in (Gala et al., 2020). Figure 1 represents the most impactful descriptors in recommending the age ranges for the Alector corpus.

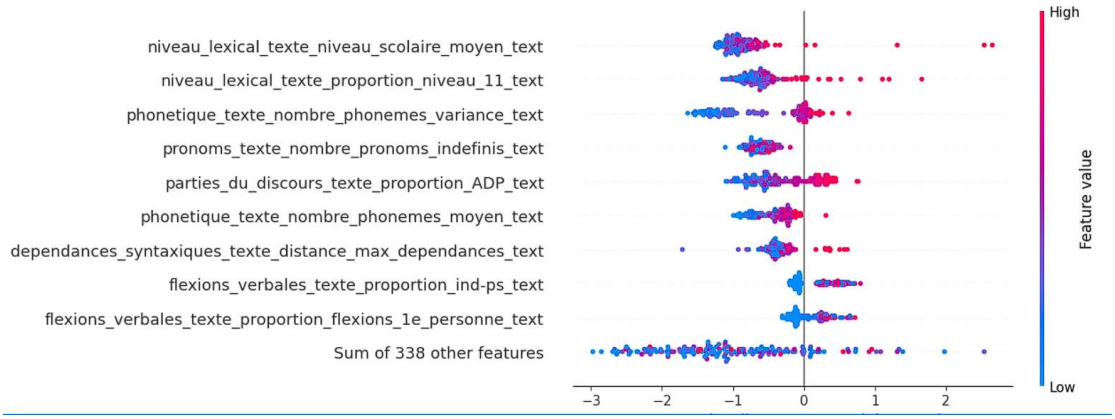

Figure 1.

Bibliography

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2020). Applying natural language processing and hierarchical machine learning approaches to text difficulty classification. International Journal of Artificial Intelligence in Education, pp. 1–34

Battistelli, D., Etienne, A., Rahman., R., & Teissèdre, C. (2022). Une chaîne de traitements pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique et psycho-linguistique. Proceedings of TALN 2022, Traitement Automatique des Langues Naturelles, pp. 236-246

Bayot, R. K., & Gonçalves, T. (2017). Age and gender classification of tweets using convolutional neural networks. International Workshop on Machine Learning, Optimization, and Big Data, pp. 337-348.

Chen, T., Guestrin, C. (2016). XGBoost : un système de boosting d'arbres évolutif. http://arxiv.org/abs/1603.02754

Chen, J., Cheng, L., Yang, X., Liang, J., Quan, B., & Li, S. (2019). Joint learning with both classification and regression models for age prediction. Journal of Physics: Conference Series

Gala, N., Tack, A., Javourey-Drevet, L., François, T., Ziegler, J.C. (2020). Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 1353-1361

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems.

Mesgar, M., & Strube, M. (2018). A neural local coherence model for text quality assessment. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4328-4339.

Rahman, R., Lecorvé, G., Etienne, A., Béchet, N., Chevelu, J. & Battistelli, D. (2020). Mama/Papa, Is this Text for Me?. Proceedingd of COLING'20 (28th International Conference on Computational Linguistics), 8-13 décembre 2020, Barcelona, Spain

Rahman, R., Lecorvé, G., & Béchet, N. (2023). Age Recommendation from Texts and Sentences for Children. Retrieved from https://arxiv.org/abs/2308.10586: https://doi.org/10.48550/arXiv.2308.10586

[1] https://texttokids.irisa.fr/
[2] https://texttokids.ortolang.fr/chain/