ANDI at SemEval-2021 Task 1: Predicting complexity in context using distributional models, behavioural norms, and lexical resources

Armand Rotaru
Independent researcher
armand.rotaru@gmail.com

Abstract

In this paper we describe our participation in the Lexical Complexity Prediction (LCP) shared task of SemEval 2021, which involved predicting subjective ratings of complexity for English single words and multi-word expressions, presented in context. Our approach relies on a combination of distributional models, both context-dependent and context-independent, together with behavioural norms and lexical resources.

1 Introduction

In our day-to-day life, outside the laboratory, we almost never come across single words or pairs of words, in isolation. Instead, such verbal stimuli are typically embedded within sentences or phrases, and our understanding of individual words and word pairs is influenced by their linguistic contexts (e.g., by disambiguating their intended meaning). Hoewever, almost all behavioural norms collected so far focus only on single words or word pairs (Johns, Jamieson, & Jones, 2020).

Therefore, the Lexical Complexity Prediction (LCP) shared task (Shardlow, Evans, Paetzold, & Zampieri, 2021), hosted at SemEval 2021, constitutes a timely and valuable contribution to the study of context-dependent semantics. The task requires competitors to predict subjective ratings of complexity for words or pairs of words, presented within sentences. As mentioned by the organisers, being able to automatically estimate contextualised complexity ratings would have several practical applications, such as detecting and simplifying portions of text that might be particularly difficult to understand for second language learners, and people with low literacy levels (e.g., as a result of suffering from a reading impairment).

The dataset for the competition is CompLex 2.0 (Shardlow, Cooper, & Zampieri, 2020; Shardlow, Evans, & Zampieri, 2021), consisting of passages from the Bible, the proceedings of the European Parliament, and biomedical journal articles. The training data covers 7,662 single words (2,574 bible, 2,512 europarl, and 2,576 biomed), and 1,517 multi-word expressions (505 bible, 498 europarl, and 514 biomed). The test data covers 917 single words (283 bible, 345 europarl, and 289 biomed), and 184 multi-word expressions (66 bible, 65 europarl, and 53 biomed).

In this paper we describe our submission to the competition, based on distributional models, both context-dependent and context-independent, as well as behavioural norms/lexical resources[1]. The best results are obtained by combining the three classes of predictors. However, the improvement in performance over using just context-independent models is small, and, in practice, might be compensated by their impressive vocabulary size and ease of use.

2 General Description

In order to predict word complexity in context, we combined information from three type of sources, namely behavioural norms/lexical resources, and distributional models. With respect to the latter, we included two distinct classes of models:

- context-independent models, which output the same vector representation for a given word, regardless of the context in which the word is encountered;

[1] https://github.com/armandrotaru/TeamAndi-LCP

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 655–660
Bangkok, Thailand (online), August 5–6, 2021.

- context-dependent models, which output a potentially different representations for a given word, as a function of the context in which the word is presented.

Our approach was very similar to that employed in (Rotaru, 2020), for predicting ratings of concreteness in context.

Firstly, we used behavioural norms collected for a wide variety of psycholinguistic factors, as well as lexical resources. More specifically, we focused on norms for concreteness (Brysbaert, Warriner, & Kuperman, 2014; Paetzold & Specia, 2016), imageability (Paetzold & Specia, 2016), semantic diversity (Hoffman, Lambon Ralph, & Rogers, 2013), age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Paetzold & Specia, 2016), familiarity (Paetzold & Specia, 2016), emotional dimensions (i.e., valence, arousal, and dominance; Mohammad, 2018), and sensorimotor dimensions (i.e., modality strengths for the tactile, auditory, olfactory, gustatory, visual, and interoceptive modalities; interaction strengths for the mouth/throat, hand/arm, foot/leg, head excluding mouth/throat, and torso effectors; Lynott, Connell, Brysbaert, Brand, & Carney, 2019). We also included complexity ratings (Maddela & Xu, 2018), lexical decision response times and accuracies (Keuleers, Lacey, & Brysbaert, 2012), contextual diversity counts (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), frequency counts (Van Heuven et al., 2014; Lin et al. 2012), prevalence counts (Brysbaert, Mandera, McCormick, & Keuleers, 2019), and CEFR word lists (Council of Europe, 2001). Nearly all these measures are correlated with word complexity.

Secondly, we employed context-independent distributional models, namely Skip-gram (Mikolov, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), and ConceptNet NumberBatch (Speer, Chin, & Havasi, 2017). Such models have been used in order to accurately predict a range of psycholinguistic variables (e.g., Hollis, Westbury, & Lefsrud, 2017; Utsumi, 2020), which suggests that they could be useful in accounting for complexity ratings.

Thirdly, we employ context-dependent distributional models, namely BERT (Devlin, Chang, Lee, & Toutanova, 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark, Luong, Le, & Manning, 2020), ALBERT (Lan et al., 2020), and DeBERTa (He, Liu, Gao, & Chen, 2020). Given that such models achieve human-level performance in various linguistic tasks (e.g., those in the GLUE benchmark; Wang et al., 2018), and that they were designed to process rich contextual information, they could be a valuable tool for predicting ratings of complexity in context.

3 System Description

We tested three groups of predictors, both in isolation and combined. The first group was obtained from comprehensive datasets of subjective ratings (concreteness, age of acquisition, etc.), task performance measures (i.e., response times and accuracies in the lexical decision tasks), as well as frequency, contextual diversity, and prevalence counts, plus CEFR word lists (see the references from the beginning of the previous section). In order to extend the coverage of the subjective ratings, we did not use the original data, but instead relied on extrapolated ratings for more than 70,000 words. The extrapolation was based on the Skip-gram, GloVe, and ConceptNet NumberBatch models, using linear regression over the concatenated vector dimensions. For the (already extrapolated) ratings from (Paetzold & Specia, 2016), as well as for the frequency, contextual diversity, and prevalence counts, we employed only the normed values, as they already have very good coverage. We also used only the original lexical decision data, given that response times and accuracies are difficult to extrapolate, and did not try to extend the CEFR word lists, due to methodological difficulties. For the single word datasets, we employed all the previously mentioned factors, whereas for the multi-word expression datasets, we only employed our own extrapolated factors.

The second group was generated from Skip-gram, GloVe, and ConceptNet NumberBatch embeddings. The vocabulary of the models was that described in the discussion above.

For the first two sources of information, and for each selected variable V (e.g., semantic diversity), we generated either four predictors, in the case of the single word datasets, or nine predictors, in the case of the multi-word expression datasets. The single word predictors consisted of $V(w)$, $V(c)$, $V(w) * V(c)$, and $\text{abs}(V(w) - V(c))$, while the multi-word expression predictors consisted of $V(w_1)$, $V(w_2)$, $V(c)$, $\text{abs}(V(w_1) - V(c))$, $\text{abs}(V(w_2) - V(c))$, $\text{abs}(V(w_1) - V(w_2))$, $V(w_1) * V(c)$, $V(w_2) * V(c)$, $V(w_1) * V(w_2)$, where:

- $V(w)$ denotes the value of V corresponding to the single word w (e.g., w = "sons"). If w is not present in our norms/models, we set $V(w)$ to the average value of V, computed over the entire vocabulary;
- $V(w_1)$ and $V(w_2)$ denote the values of V corresponding to the words w_1 and w_2 (e.g., w_1 = "skillful", w_2 = "workman"), that make up the multi-word expression w_1 w_2 (i.e., w_1 w_2 = "skillful workman"). As before, if w_1 and/or w_2 are not present in our norms/models, we set $V(w_1)$ and/or $V(w_2)$ to the average value of V, computed over the entire vocabulary;
- $V(c)$ denotes the value of V corresponding to the context c in which the single word w, or multi-word expression w_1 w_2, are encountered (e.g., w = "sons", c = "The ____ of Perez: Hezron, and Hamul."; or w_1 w_2 = "skillful workman", c = "He made it the work of a ____."). Computing this value involves calculating the average $V(c) = \frac{\sum_{i=1}^{N} V(c_i)}{N}$, where $V(c_i)$ is the value of V corresponding to the i-th context word, calculated as described previously, and N is the number of context words.

These predictors allowed us to include both the individual contributions of the single word w, or the multi-word expression w_1 w_2, and the context c, as well as certain interactions between the former and the latter.

The third group was derived from the BERT, RoBERTa, ELECTRA, ALBERT, and DeBERTa models. We used the standard (base) versions of each model (i.e., without task-specific fine-tuning), as described in the original papers, with the exception of ELECTRA, where we employed the small, base, and large versions of the model. The implementations of the models were all obtained from the Hugging Face repository (Wolf et al., 2020). The predictors consisted only of the activations for the single word w, or the multi-word expression w_1 w_2, averaged over the last four hidden layers.

To predict ratings of complexity in context, we employed ridge regression (lambda = 3000), for the single word dataset, and a combination of ridge regression (lambda = 1250) and gradient-boosted decision trees, for the multi-word expression dataset, after zero centering all the aforementioned predictors.

4 Results and Discussion

The results for English are shown Figure 1, for various sets of predictors and regularization strengths. For reasons of space, we only present the results for ridge regression, but note that similar patterns of performance are obtained for gradient-boosted decision trees and other types of models, such as shallow neural networks. Results are averaged over 10 rounds of 10-fold cross-validation, using only the training dataset.

The results indicate that context-independent models (Fig. 1b) outperform behavioural norms (Fig. 1a), and context-dependent models (Fig. 1c-f). A likely reason for the superiority of context-independent models over context-dependent models is the fact that the former were trained on huge corpora (i.e., 100-840 billion tokens), while the latter were trained on considerably smaller corpora (i.e., 3-33 billion tokens). However, in spite of this significant training disadvantage, context-dependent models produce competitive levels of performance, a finding which can likely be attributed to several factors, such as the highly non-linear integration of contextual information, the use of self-attention mechanisms, and that of more sophisticated learning objectives.

Combining the three classes of predictors produces a relatively small improvement in predictive performance, as compared to relying on any single class. This reflects a very high degree of redundancy between the complexity-related information present in the three types of predictors.

Interestingly, even for the largest set of predictors, consisting of 13,400 variables per 1,517 data points, the degree of regularization does not appear to matter much, indicating little overfitting.

Finally, there is a small, but systematic difference in performance between single words and multi-word expressions, in favour of the latter, even though the training set for single word stimuli is roughly five times larger than that for multi-word stimuli. A potential explanation for this finding might be that the individual variability in meaning for multi-word expressions is smaller than that for single words, given that expressions should be more informative than single words, in virtue of their length (i.e., two words vs one word).

Within the competition, our models ranked 4th (r = .78, ρ = .73, MAE = .064), in the single word sub-task, and 6th (r = .85, ρ = .84, MAE = .067), in the multi-word expression sub-task.

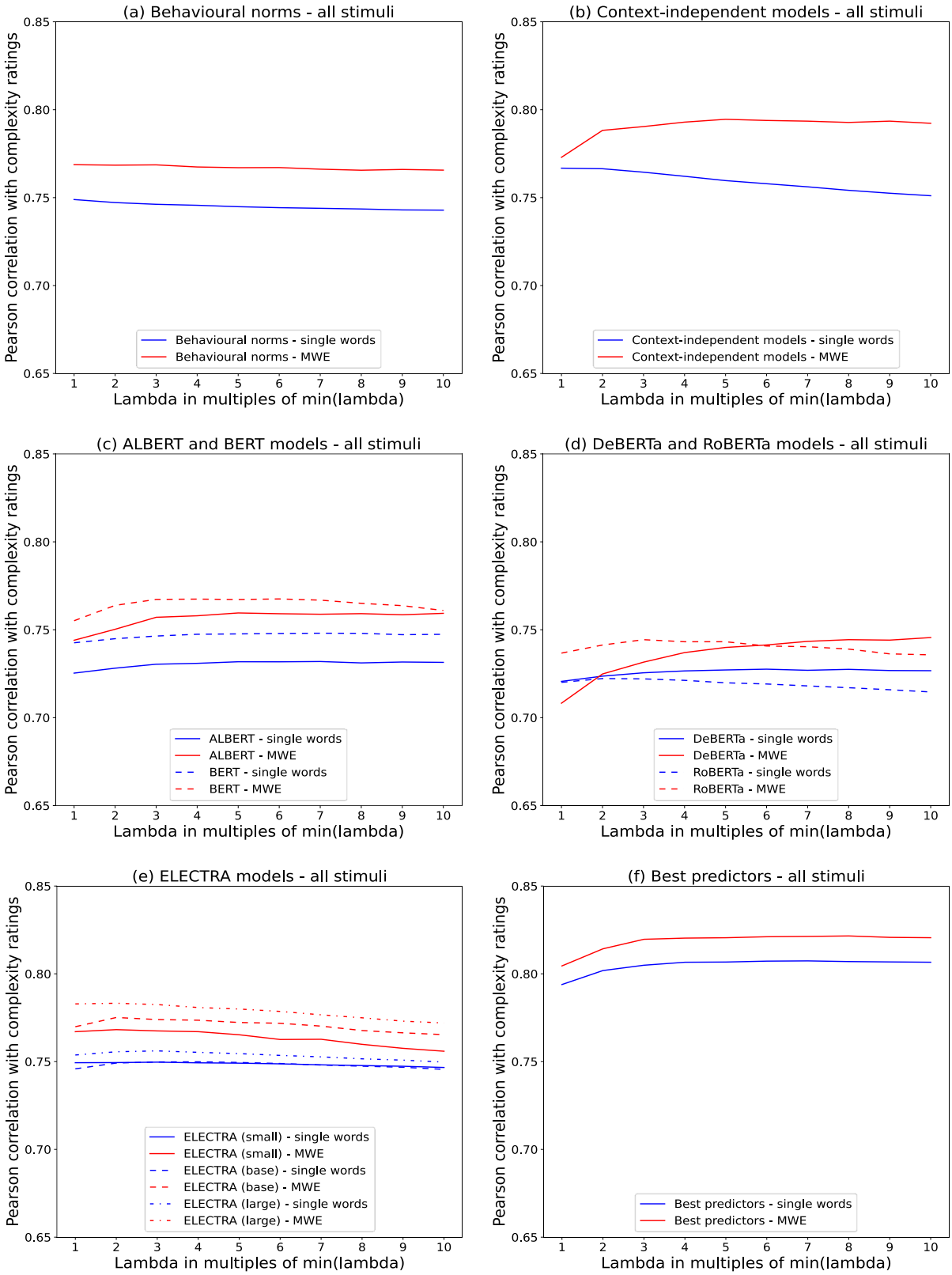

Figure 1. Pearson correlations between predicted and actual complexity ratings, for various groups of predictors and regularization strengths (i.e., values of lambda). For single words, in subfigures (a)-(e), min(lambda) = 100, while in subfigure (f), min(lambda) = 400. For multi-word expressions, in subfigures (a)-(e), min(lambda) = 50, while in subfigure (f), min(lambda) = 200.

5 Conclusions

Our results suggest that several approaches can be quite successfully employed in order to predict ratings of complexity in context, for both single words and multi-word expressions. In terms of performance, the best predictors are those derived from context-independent models (e.g., Skip-gram), but relatively good results can be obtained also by using context-dependent models (e.g., BERT) and behavioural norms (e.g., subjective ratings of familiarity). Moreover, given that their vocabulary covers a remarkable number of words (i.e., more than 500 thousand, for each of the Skip-gram, GloVe, and ConceptNet NumberBatch models), and that they are very easy to use off-the-shelf, context-independent models represent a particularly promising approach to predicting ratings of complexity in context.

References

Marc Brysbaert, Amy B. Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904-911.

Marc Brysbaert, Paweł Mandera, Samantha F. McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51(2): 467-479.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the ICLR*. Pages 1-18.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

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 NAACL-HLT*. Association for Computational Linguistics, pages 4171-4186.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv preprint arXiv:2006.03654.

Paul Hoffman, Matthew A. Lambon Ralph, and Timothy T. Rogers. 2013. Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45(3):718-730.

Geoff Hollis, Chris Westbury, and Lianne Lefsrud. 2017. Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, 70(8):1603-1619.

Brendan T. Johns, Randall K. Jamieson, and Michael N. Jones. 2020. The continued importance of theory: Lessons from big data approaches to language and cognition. In *Big data methods for psychological research: New horizons and challenges*. APA, pages 277-295.

Emmanuel Keuleers, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44(1):287-304.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the ICLR*. Pages 1-17.

Yuri Lin, Jean-Baptiste Michel, Erez L. Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings of the 50th Annual Meeting of the ACL*. Association for Computational Linguistics, pages 169-174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint:1907.11692.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52:1-21.

Mounica Maddela and Wei Xu. (2018). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 3749-3760.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the Workshop at the ICLR*. Pages 1-12.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the ACL - Long Papers*. Association for Computational Linguistics, pages 174-184.

Gustavo H. Paetzold and Lucia Specia. 2016. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 435-440.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1532-1543.

Armand Rotaru. 2020. ANDI @ CONCRETEXT: Predicting concreteness in context for English and Italian using distributional models and behavioural norms. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020).* Online. CEUR.org.

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)*. European Language Resources Association, pages 57-62.

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

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI*. AAAI Press, pages 4444-4451.

Akira Utsumi. 2020. Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cognitive Science*, 44(6):e12844.

Walter J.B. 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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP*. Association for Computational Linguistics, pages 353-355.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pages 38-45.