PolyU-CBS at TSAR-2022 Shared Task: A Simple, Rank-Based Method for Complex Word Substitution in Two Steps

Emmanuele Chersoni and **Yu-Yin Hsu**
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong, China
emmanuelechersoni@gmail.com, yu-yin.hsu@polyu.edu.hk

Abstract

In this paper, we describe the system we present at the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022) regarding the shared task on Lexical Simplification for English, Portuguese, and Spanish. We proposed an unsupervised approach in two steps: First, we used a masked language model with word masking for each language to extract possible candidates for the replacement of a difficult word; second, we ranked the candidates according to three different Transformer-based metrics. Finally, we determined our list of candidates based on the lowest average rank across different metrics. The results show that our method, based on two simple steps and rankings, can effectively improve the scores among datasets for the task of lexical simplification.

1 Introduction

The notion of *linguistic complexity* has been widely debated in both theoretical and computational linguistics, and has been interpreted very differently depending on the discipline. Specifically, in the field of natural language processing (NLP), complexity has often been associated with the difficulties that language users encounter while processing concrete linguistic productions (e.g., sentences, utterances, etc.) (Blache, 2011; Chersoni et al., 2016, 2017, 2021; Sarti et al., 2021; Iavarone et al., 2021), with research focusing on applications that aim to simplify challenging texts and to make them more easily readable for a wider variety of users (North et al., 2022b).

Previously, NLP shared tasks focused on the problem of identifying a complex word in a sentence, or assigning a difficulty score to it (Yimam et al., 2018; Shardlow et al., 2021). The TSAR-2022 shared task (Saggion et al., 2022) instead focused on the next step; that is, how to find simpler words as replacement candidates for a given target word in a multilingual setting. Consequently, the task can be seen as similar to lexical substitution in context (McCarthy and Navigli, 2009).

In this paper, we describe our contribution to the TSAR-2022 shared task, which is a system for English, Portuguese, and Spanish that i) generates replacement candidates for a given word via masked language modeling, and ii) assigns scores to the candidates by averaging the ranks assigned by different Transformer-based metrics.

2 Related Work

The goal of the previous shared tasks regarding lexical complexity was to identify complex words in a sentence context, and complexity was defined as a binary variable (Paetzold and Specia, 2016b; Yimam et al., 2018). However, these tasks were oversimplified because there is no clear-cut choice in many contexts, and human annotators prefer to assign a score based on a continuous scale of difficulty. Shardlow et al. (2020) introduced the CompLex corpus, a gold-standard benchmark for lexical complexity in English, in which words and multiword expressions are extracted from different text genres (legal, religious, and biomedical genres) and are annotated with continuous scores that reflect their difficulty in the sentence context. The same corpus was then used as the source material for the SemEval-2021 shared task regarding lexical complexity in context (Shardlow et al., 2021).

The estimation of lexical complexity is only one component in the lexical simplification pipeline, which also involves generating candidates for substitution, ranking them, and assessing their degree of fitness in the given sentence context. Datasets focusing on the latter parts of the pipeline have been published for English (Specia et al., 2012; Horn et al., 2014; Paetzold and Specia, 2016a; Štajner et al., 2022), Japanese (Kajiwara and Yamamoto, 2015; Hading et al., 2016), Portuguese (Hartmann and Aluísio, 2020; North et al., 2022a; Štajner et al., 2022), French (Rolin et al., 2021), Spanish (Alar-

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 225 - 230

Language	Sentence	Target	Substitutes
English (EN)	Brevard County was the scene of six homicides in 2011, Goodyear said.	homicides	murders deaths killings
Portuguese (PT)	o nosso é brasileiro colorido é um menino alegre com pontos de melancolia	melancolia	tristeza tédio abatimento
Spanish (ES)	Antes de aquello, el estadio albergaba una capacidad para más de 130.000 espectadores.	albergaba	alojaba tiene aloja

Table 1: Dataset examples for each of the three languages.

con, 2021; Ferrés and Saggion, 2022; Štajner et al., 2022), and Chinese (Qiang et al., 2021).

The current state-of-the-art system for English, LSBert, was introduced by Qiang et al. (2020). The system first generates a list of possible replacement candidates via the masked language modeling function of BERT (Devlin et al., 2019) by being fed the original sentence concatenated with a copy of the sentence in which the original word has been masked. The system then performs a re-ranking using different features, e.g. frequency, vector-based semantic similarity, and/or language model probability. Studies using LSBert (Przybyła and Shardlow, 2020; Štajner et al., 2022) have shown that the approach could easily be adapted to other languages and still achieve state-of-the-art results.

3 Experimental Settings

3.1 Datasets

The shared task organizers provided a testing dataset (Štajner et al., 2022) with a combined number of 1115 instances: 373 for English, 374 for Portuguese, and 368 for Spanish. Each instance consisted of a sentence, a target word, and a list with a variable number of gold replacement words, all obtained from human native speakers on Amazon Mechanical Turk. Each instance was annotated by 25 different annotators, and each annotator had to simplify the sentence by proposing a simpler candidate word for substitution. An example for each target language is displayed in Table 1.

3.2 Methodology

3.2.1 Candidate Generation

For each of the three target languages, we masked each target word in the dataset instances and used a masked language model – a variant of the BERT Base model (Devlin et al., 2019) – to generate a list of candidate words (the original word itself was filtered out). For English, we simply used the original BERT Base;[1] for Portuguese, we used the BERT Base BERTimbau model by Souza et al. (2020);[2] for Spanish, we used the BETO model by Canete et al. (2020).[3] For our experiments, the number n of generated candidates was used as a system parameter, and was fixed at $n = 30$. Importantly, for each candidate word we saved the $rank$; that is, the position that the word occupies in the list of candidates sorted by decreasing probability score. We refer to this method, before any re-ranking step, as the **Base** and used it as a baseline method.

3.2.2 Candidate Re-Ranking

Using the n candidate words identified in the candidate generation step, we extracted three Transformer-based metrics for re-ranking. The idea behind our approach is that words that achieve higher scores and lower rankings for multiple metrics are strong candidates for replacement.

We considered three metrics, which we extracted via the `minicons` library (Misra, 2022):

- *Sentence probability via autoregressive language modeling*. For each item, we replaced the target word with a candidate substitute word, and computed a probability for the whole sentence via a variant of the GPT2 model (Radford et al., 2019). For English, we used the original GPT2-Base;[4] for Portuguese, the GPorTuguese-2 Small (Guillou, 2020);[5]

[1] https://huggingface.co/bert-base-uncased

[2] https://huggingface.co/neuralmind/bert-base-portuguese-cased

[3] https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

[4] https://huggingface.co/gpt2

[5] https://huggingface.co/pierreguillou/gpt2-small-portuguese

Method	Acc@1	Acc(1,2,3)@Top1	Pot(3,5,10)	MAP(3,5,10)
Base	0.27	0.12 / 0.19 / 0.22	0.49 / 0.57 / 0.68	0.17 / 0.13 / 0.08
Base + LMProb *	0.32	0.14 / 0.20 / 0.26	0.51 / 0.60 / 0.71	0.19 / 0.15 / 0.09
Base + PLL	0.29	0.12 / 0.17 / 0.22	0.5 / 0.6 / 0.72	0.18 / 0.14 / 0.08
Base + cosSim *	0.43	0.2 / 0.28 / 0.33	0.61 /0.7 / 0.77	0.27 / 0.2 / 0.11
Base + All *	0.4	0.18 / 0.26 / 0.3	0.59 / 0.68 / 0.75	0.25 / 0.18 / 0.11
TUNER	0.34	0.14 / 0.17 / 0.18	0.43 / 0.44 / 0.44	0.17 / 0.1 / 0.05
LSBert	**0.6**	**0.3 / 0.44 / 0.53**	**0.82 / 0.87 / 0.94**	**0.40 / 0.29 / 0.17**

Table 2: Scores for the English dataset. * indicates the systems submitted to the shared task.

and for Spanish, a GPT2 Base model trained on the BETO corpus (Canete et al., 2020).[6]

- *Sentence probability via masked language modeling*. Similar to the previous metric, we computed the probability of the sentence via estimating the pseudo-log-likelihood (PLL) with a masked language model (the scores were obtained by masking the tokens one-by-one) (Salazar et al., 2020). For this metric, we adopted the same versions of BERT Base used in the step of candidate generation.

- *Contextualized embedding similarity*. By always using the same BERT Base models, we measured the cosine similarity of i) the contextualized embedding of the target word in the context of the original sentence, and ii) the contextualized embedding of each candidate word after replacing the target word in the original sentence.

$$score(w) = \frac{rank_{Base}(w) + rank_{metric}(w)}{2} \quad (1)$$

After computing the scores for each of the three metrics in our pool of n candidates, we sorted them to obtain their respective rankings. We call these rankings, respectively, *LMProb*, *PLL* and *cosSim*. Then, for each candidate word w, we computed its score by averaging the rank in the **Base** model and the rank in one of the metrics (see Equation 1). This resulted in three different scores: 1) **Base + LMProb**; 2) **Base + PLL**; and 3) **Base + cosSim**. We then computed one last score, which averaged the ranks of the four rankings together for each candidate word. We call this score **Base + All**. The scores of the candidate words are finally sorted in ascending order (the ones with the lowest ranks are the top candidates for replacement).

[6] https://huggingface.co/mrm8488/spanish-gpt2

3.3 Baselines and State-of-the-Art

We presented the scores for a simple baseline method, based on the mere candidate generation by a BERT masked language model, without any further re-ranking (**Base**). Moreover, the scores for two state-of-the-art systems were provided by the shared task organizers for comparison:

- **TUNER**, an unsupervised system introduced by Ferrés et al. (2017) for Spanish, and further adapted to English and Portuguese. The system relies on the identification of a list of candidate synonyms via a word sense disambiguation algorithm and a distributional thesaurus.[7] Candidates are then re-ranked based on their frequencies in the Wikipedia of each language. Finally, a morphological generator component ensures that the correct form of the word is selected for the final replacement;

- The above-mentioned **LSBert** system (Qiang et al., 2020), with its adaptations to Spanish and Portuguese.

3.4 Evaluation

Evaluation metrics for lexical simplification were introduced by Paetzold and Specia (2016a):

- *Accuracy (Acc)*: $Acc@1$ is the ratio of instances for which the top substitute is in the gold standard, regardless of the order, and it is the main metric for ranking the shared task systems; $AccK$ measures instead the ratio of instances for which at least one of the top K predicted candidates matches the most frequently suggested candidate synonym in the gold standard (we made our system return up to 10 candidates per instance);

[7] Both tools rely on the Freeling text analysis tool (Padró and Stanilovsky, 2012), available at: https://nlp.lsi.upc.edu/freeling/index.php/node/1.

Method	Acc@1	Acc(1,2,3)@Top1	Pot(3,5,10)	MAP(3,5,10)
Base	0.23	0.1 / 0.12 / 0.15	0.34 / 0.39 / 0.49	0.12 / 0.08 / 0.05
Base + LMProb *	0.22	0.09 / 0.12 / 0.15	0.33 / 0.38 / 0.49	0.11 / 0.08 / 0.05
Base + PLL	0.22	0.09 / 0.13 / 0.14	0.34 / 0.4 / 0.48	0.12 / 0.08 / 0.05
Base + cosSim *	0.32	0.14 / 0.19 / 0.21	0.45 / 0.51 / 0.57	0.17 / 0.12 / 0.07
Base + All *	0.28	0.11 / 0.14 / 0.17	0.4 / 0.47 / 0.55	0.15 / 0.1 / 0.06
TUNER	0.22	0.13 / 0.16 / 0.16	0.27 / 0.27 / 0.27	0.1 / 0.06 / 0.03
LSBert	**0.32**	**0.16 / 0.23 / 0.28**	**0.49 / 0.58 / 0.67**	**0.19 / 0.13 / 0.07**

Table 3: Scores for the Portuguese dataset. * indicates the systems submitted to the shared task.

Method	Acc@1	Acc(1,2,3)@Top1	Pot(3,5,10)	MAP(3,5,10)
Base	0.24	0.1 / 0.14 / 0.18	0.45 / 0.53 / 0.62	0.15 / 0.11 / 0.06
Base + LMProb *	0.2	0.08 / 0.13 / 0.17	0.41 / 0.5 / 0.64	0.14 / 0.1 / 0.06
Base + PLL	0.23	0.08 / 0.15 / 0.2	0.44 / 0.54 / 0.64	0.16 / 0.11 / 0.06
Base + cosSim *	**0.36**	**0.16 / 0.2 / 0.23**	**0.52 / 0.6 / 0.68**	**0.2 / 0.14 / 0.08**
Base + All *	0.28	0.11 / 0.18 / 0.22	0.5 / 0.6 / 0.68	0.18 / 0.13 / 0.07
TUNER	0.12	0.06 / 0.08 / 0.08	0.14 / 0.14 / 0.15	0.06 / 0.03 / 0.02
LSBert	0.28	0.09 / 0.14 / 0.18	0.49 / 0.61 / 0.74	0.19 / 0.13 / 0.07

Table 4: Scores for the Spanish dataset. * indicates the systems submitted to the shared task.

- *Potential (Pot)*: the ratio of instances for which at least one of the generated candidates is present in the gold standard.
- *Mean Average Precision (MAP)*: a commonly-used metric in information retrieval, which assesses how many of the predicted candidates are relevant (i.e., how many of them are present in the gold standard annotations).

In the official results, the metrics are computed based on different values of K: for Accuracy, $K = 1, 2, 3$, while for Potential and MAP, $K = 3, 5, 10$.

4 Results and Conclusion

The results for English, Portuguese and Spanish can be seen, respectively, in Table 2, 3 and 4. On the basis of preliminary results on the trial dataset, we submitted the scores for Base + All, Base + LMProb, and Base + cosSim in all the three language tracks. At a glance, it can be seen that the combination of the Base ranking with the ranking based on cosine similarity is the only one that consistently improves over the baseline performance. A possible reason is that the initial selection of the candidates is already based on a Transformer language model, so it could be the case that the information coming from the language model-based rankings is redundant, or tend to suggest the same subset of candidates. On the other hand, the cosine metric between the contextualized embeddings is assessing a paradigmatic type of similarity between the target and the candidate word: this is not necessarily taken into account by the other metrics, which are more focused on the syntagmatic axis.

Our method relying on Base + cosSim, which was submitted as PolyU-CBS3, was the one reporting the best scores on all the three datasets (15th overall on English, 5th on Spanish, 3rd on Portuguese). It is noticeable that our methods always outperform TUNER on the metrics of Potential and MAP. The LSBert is the best performing method on English and Portuguese datasets, although our Base + cosSim is a close match to the latter. Finally, Base + cosSim outperforms both TUNER and LSBert on Spanish. We take the results as a preliminary evidence that our method, based on two simple steps and ranking, can be highly effective for the task of lexical simplification. A possible way to further improve the methodology will be to introduce different methods of extracting candidate words. In our preliminary experiments, we found that a similarity ranking based on traditional, static embedding model alone can lead to improvements of the performance on English. However, for languages with a richer morphology like the Romance ones, a morphological adapter would be needed to generate the form that best fits the target sentence. Another possible direction could be using a generative model treating the task as a text-to-text problem (Raffel et al., 2020), which could be fine-tuned on supervised lexical substitution data and combined with a frequency filter to ensure that the proposed replacement is actually a simpler word.

Acknowledgements

This study was supported by the Startup fund (1-BD8S) by the Hong Kong Polytechnic University.

References

Rodrigo Alarcon. 2021. Dataset of Sentences Annotated With Complex Words and Their Synonyms to Support Lexical Simplification. *Mendeley Data*.

Philippe Blache. 2011. Evaluating Language Complexity in Context: New Parameters for a Constraint-based Model. In *Proceedings of the International Workshop on Constraints and Language Processing*.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Proceedings of the ICLR Workshop on Practical Machine Learning for Developing Countries*.

Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a Distributional Model of Semantic Complexity. In *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity*.

Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical Metonymy in a Distributional Model of Sentence Comprehension. In *Proceedings of *SEM*.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021. Not All Arguments Are Processed Equally: A Distributional Model of Argument Complexity. *Language Resources and Evaluation*, 55(4):873–900.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A Dataset for Lexical Simplification in Spanish. In *Proceedings of LREC*.

Daniel Ferrés, Horacio Saggion, and Xavier Gómez Guinovart. 2017. An Adaptable Lexical Simplification Architecture for Major Ibero-Romance Languages. In *Proceedings of the EMNLP Workshop on Building Linguistically Generalizable NLP Systems*.

Pierre Guillou. 2020. GPorTuguese-2 (Portuguese GPT-2 Small): A Language Model for Portuguese Text Generation (and More NLP Tasks...).

Muhaimin Hading, Yuji Matsumoto, and Maki Sakamoto. 2016. Japanese Lexical Simplification for Non-native Speakers. In *Proceedings of the COLING Workshop on Natural Language Processing Techniques for Educational Applications*.

Nathan Siegle Hartmann and Sandra Maria Aluísio. 2020. Adaptaçao Lexical Automática em Textos Informativos do Português Brasileiro para o Ensino Fundamental. *Linguamática*, 12(2):3–27.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of ACL*.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence Complexity in Context. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*.

Diana McCarthy and Roberto Navigli. 2009. The English Lexical Substitution Task. *Language Resources and Evaluation*, 43(2):139–159.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022a. ALEXSIS-PT: A New Resource for Portuguese Lexical Simplification. *arXiv preprint arXiv:2209.09034*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2022b. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys (CSUR)*.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards Wider Multilinguality. In *Proceedings of LREC*.

Gustavo Paetzold and Lucia Specia. 2016a. Benchmarking Lexical Simplification Systems. In *Proceedings of LREC*.

Gustavo Paetzold and Lucia Specia. 2016b. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of SemEval*.

Piotr Przybyła and Matthew Shardlow. 2020. Multi-Word Lexical Simplification. In *Proceedings of COLING*.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. LSBert: A Simple Framework for Lexical Simplification. *arXiv preprint arXiv:2006.14939*.

Jipeng Qiang, Xinyu Lu, Yun Li, Yunhao Yuan, and Xindong Wu. 2021. Chinese Lexical Simplification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 29:1819–1828.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-text

Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Eva Rolin, Quentin Langlois, Patrick Watrin, and Thomas François. 2021. FrenLyS: A Tool for the Automatic Simplification of French General Language Texts. In *Proceedings of RANLP*.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification. In *Proceedings of the EMNLP Workshop on Text Simplification, Accessibility, and Readability*.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of ACL*.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That Looks Hard: Characterizing Linguistic Complexity in Humans and Language Models. In *Proceedings of the NAACL Workshop on Cognitive Modeling and 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 LREC Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of SemEval*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 Task 1: English Lexical Simplification. In *Proceedings of SemEval*.

Sanja Štajner, Daniel FerrĂŹes, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical Simplification Benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*, 5.

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. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.